1

Introduction to Multilevel Analysis

Social research regularly involves problems that investigate the relationship between individuals and society. The general concept is that individuals interact with the social contexts to which they belong, that individual persons are influenced by the social groups or contexts to which they belong, and that those groups are in turn influenced by the individuals who make up that group. The individuals and the social groups are conceptualized as a hierarchical system of individuals nested within groups, with individuals and groups defined at separate levels of this hierarchical system. Naturally, such systems can be observed at different hierarchical levels, and variables may be defined at each level. This leads to research into the relationships between variables characterizing individuals and variables characterizing groups, a kind of research that is generally referred to as *multilevel research*.

In multilevel research, the data structure in the population is hierarchical, and the sample data are a sample from this hierarchical population. Thus, in educational research, the population consists of schools and pupils within these schools, and the sampling procedure often proceeds in two stages: First, we take a sample of schools, and next we take a sample of pupils within each school. Of course, in real research one may have a convenience sample at either level, or one may decide not to sample pupils but to study all available pupils in the sample of schools. Nevertheless, one should keep firmly in mind that the central statistical model in multilevel analysis is one of successive sampling from each level of a hierarchical population.

In this example, pupils are *nested* within schools. Other examples are cross-national studies where the individuals are nested within their national units, organizational research with individuals nested within departments within organizations, family research with family members within families, and methodological research into interviewer effects with respondents nested within interviewers. Less obvious applications of multilevel models are longitudinal research and growth curve research, where a series of several distinct observations are viewed as nested within individuals, and meta-analysis where the subjects are nested within different studies. For simplicity, this book describes the multilevel models mostly in terms of individuals nested within groups, but note that the models apply to a much larger class of analysis problems.

1.1 AGGREGATION AND DISAGGREGATION

In multilevel research, variables can be defined at any level of the hierarchy. Some of these variables may be measured directly at their 'own' natural level; for example, at the school level we may measure school size and denomination, and at the pupil level intelligence and school success. In addition, we may move variables from one level to another by aggregation or disaggregation. Aggregation means that the variables at a lower level are moved to a higher level, for instance by assigning to the schools the school mean of the pupils' intelligence scores. Disaggregation means moving variables to a lower level, for instance by assigning to all pupils in the schools a variable that indicates the denomination of the school they belong to.

The lowest level (level 1) is usually defined by the individuals. However, this is not always the case. Galtung (1969), for instance, defines roles within individuals as the lowest level, and in longitudinal designs, repeated measures within individuals are the lowest level.

At each level in the hierarchy, we may have several types of variables. The distinctions made in the following are based on the typology offered by Lazarsfeld and Menzel (1961), with some simplifications. In our typology, we distinguish between *global*, *structural* and *contextual* variables.

Global variables are variables that refer only to the level at which they are defined, without reference to other units or levels. A pupil's intelligence or gender would be a global variable at the pupil level. School size would be a global variable at the school level. A global variable is measured at the level at which that variable actually exists.

Structural variables are operationalized by referring to the sub-units at a lower level. They are constructed from variables at a lower level, for example, in defining the school variable 'mean intelligence' as the mean of the intelligence scores of the pupils in that school. Using the mean of a lower-level variable as an explanatory variable at a higher level is a common procedure in multilevel analysis. Other functions of the lower-level variables are less common, but may also be valuable. For instance, using the standard deviation of a lower-level variable as an explanatory variable at a higher level could be used to test hypotheses about the effect of group heterogeneity on the outcome variable. Klein and Kozlowski (2000) refer to such variables as configural variables, and emphasize the importance of capturing the pattern of individual variation in a group. Their examples also emphasize the use of other functions than the mean of individual scores to reflect group characteristics.

It is clear that constructing a structural variable from the lower-level data involves aggregation. *Contextual* variables, on the other hand, refer to the super-units; all units at the lower level receive the value of a variable for the super-unit to which they belong at the higher level. For instance, we can assign to all pupils in a school the

school size, or the mean intelligence, as a pupil-level variable. This is called *disaggregation*; data on higher-level units are disaggregated into data on a larger number of lower-level units. The resulting variable is called a *contextual* variable, because it refers to the higher-level context of the units we are investigating.

In order to analyze multilevel models, it is not important to assign each variable to its proper place in the typology. The benefit of the scheme is conceptual; it makes clear to which level a measurement properly belongs. Historically, multilevel problems have led to analysis approaches that moved all variables by aggregation or disaggregation to one single level of interest followed by an ordinary multiple regression, analysis of variance, or some other 'standard' analysis method. However, analyzing variables from different levels at one single common level is inadequate, and leads to two distinct types of problems.

The first problem is statistical. If data are aggregated, the result is that different data values from many sub-units are combined into fewer values for fewer higher-level units. As a result, much information is lost, and the statistical analysis loses power. On the other hand, if data are disaggregated, the result is that a few data values from a small number of super-units are 'blown up' into many more values for a much larger number of sub-units. Ordinary statistical tests treat all these disaggregated data values as independent information from the much larger sample of sub-units. The proper sample size for these variables is of course the number of higher-level units. Using the larger number of disaggregated cases for the sample size leads to significance tests that reject the null-hypothesis far more often than the nominal alpha level suggests. In other words: investigators come up with many 'significant' results that are totally spurious.

The second problem is conceptual. If the analyst is not very careful in the interpretation of the results, s/he may commit the fallacy of the wrong level, which consists of analyzing the data at one level, and formulating conclusions at another level. Probably the best-known fallacy is the *ecological fallacy*, which is interpreting aggregated data at the individual level. It is also known as the 'Robinson effect' after Robinson (1950). Robinson presents aggregated data describing the relationship between the percentage of blacks and the illiteracy level in nine geographic regions in 1930. The *ecological correlation*, that is, the correlation between the aggregated variables at the region level, is 0.95. In contrast, the individual-level correlation between these global variables is 0.20. Robinson concludes that in practice an ecological correlation is almost certainly not equal to its corresponding individual-level correlation. For a statistical explanation, see Robinson (1950) or Kreft and de Leeuw (1987). Formulating inferences at a higher level based on analyses performed at a lower level is just as misleading. This fallacy is known as the *atomistic fallacy*. A related but different fallacy is known as 'Simpson's paradox' (see Lindley & Novick, 1981). Simpson's paradox refers to the problem that completely erroneous conclusions may be drawn if grouped data, drawn from heterogeneous populations, are collapsed and analyzed as if they

came from a single homogeneous population. An extensive typology of such fallacies is given by Alker (1969). When aggregated data are the only available data, King (1997) presents some procedures that make it possible to estimate the corresponding individual relationships without committing an ecological fallacy.

A better way to look at multilevel data is to realize that there is not one 'proper' level at which the data should be analyzed. Rather, all levels present in the data are important in their own way. This becomes clear when we investigate cross-level hypotheses, or *multilevel* problems. A multilevel problem is a problem that concerns the relationships between variables that are measured at a number of different hierarchical levels. For example, a common question is how a number of individual and group variables influence one single individual outcome variable. Typically, some of the higher-level explanatory variables may be the aggregated group means of lower-level individual variables. The goal of the analysis is to determine the direct effect of individual- and group-level explanatory variables, and to determine if the explanatory variables at the group level serve as moderators of individual-level relationships. If group-level variables moderate lower-level relationships, this shows up as a statistical interaction between explanatory variables from different levels. In the past, such data were usually analyzed using conventional multiple regression analysis with one dependent variable at the lowest (individual) level and a collection of explanatory variables from all available levels (see Boyd & Iversen, 1979; van den Eeden & Hüttner, 1982). Since this approach analyzes all available data at one single level, it suffers from all of the conceptual and statistical problems mentioned above.

1.2 WHY DO WE NEED SPECIAL MULTILEVEL ANALYSIS TECHNIQUES?

A multilevel problem concerns a population with a hierarchical structure. A sample from such a population can be described as a multistage sample: First, we take a sample of units from the higher level (e.g., schools), and next we sample the sub-units from the available units (e.g., we sample pupils from the schools). In such samples, the individual observations are in general not completely independent. For instance, pupils in the same school tend to be similar to each other, because of selection processes (for instance, some schools may attract pupils from higher social economic status (SES) levels, while others attract lower SES pupils) and because of the common history the pupils share by going to the same school. As a result, the average correlation (expressed in the so-called *intraclass correlation*) between variables measured on pupils from the same school will be higher than the average correlation between variables measured on pupils from different schools. Standard statistical tests lean heavily on the assumption of independence of the observations. If this assumption is violated (and in multilevel

data this is almost always the case) the estimates of the standard errors of conventional statistical tests are much too small, and this results in many spuriously 'significant' results. The effect is generally *not* negligible, as small dependencies in combination with large group sizes still result in large biases in the standard errors. The strong biases that may be the effect of violation of the assumption of independent observations made in standard statistical tests have been known for a long time (Walsh, 1947) and are still a very important assumption to check in statistical analyses (Stevens, 2009).

The problem of dependencies between individual observations also occurs in survey research, if the sample is not taken at random but cluster sampling from geographical areas is used instead. For similar reasons as in the school example given above, respondents from the same geographical area will be more similar to each other than are respondents from different geographical areas. This leads again to estimates for standard errors that are too small and produce spurious 'significant' results. In survey research, this effect of cluster sampling is well known (see Kish, 1965, 1987). It is called a 'design effect', and various methods are used to deal with it. A convenient correction procedure is to compute the standard errors by ordinary analysis methods, estimate the intraclass correlation between respondents within clusters, and finally employ a correction formula to the standard errors. A correction described by Kish (1965, p. 259) corrects the sampling variance using $v_{eff} = v(1 + (n_{clus} - 1)\rho)$, where v_{eff} is the effective sampling variance, v is the sampling variance calculated by standard methods assuming simple random sampling, n_{clus} is the cluster size, and ρ is the intraclass correlation. The corrected standard error is then equal to the square root of the effective sampling variance The intraclass correlation can be estimated using the between and within mean square from a one-way analysis of variance with the groups as a factor:

$$\rho = (MS_B - MS_W)/(MS_B + (n_{clus} - 1)MS_W).$$

The formula assumes equal group sizes, which is not always realistic. Chapter 2 presents a multilevel model that estimates the intraclass correlation (ICC) without assuming equal group sizes. A variation of the Kish formula computes the effective sample size in two-stage cluster sampling as $n_{eff} = n/[1 + (n_{clus} - 1)\,\rho]$, where n is the total sample size and n_{eff} is the effective sample size. Using this formula, we can simply calculate the effective sample size for different situations, and use weighting to correct the sample size determined by traditional software.[1] For instance, suppose that we take a sample of 10 classes, each with 20 pupils. This comes to a total sample size of 200, which is

[1] The formulas given here apply to two-stage cluster sampling. Other sampling schemes, such as stratified sampling, require different formulas. See Kish (1965, 1987) for details. The symbol ρ (the Greek letter rho) was introduced by Kish (1965, p. 161) who called it *roh* for 'rate of homogeneity'.

reasonable. Let us further suppose that we are interested in a variable for which the intraclass correlation ρ is .10. This seems a rather low intraclass correlation. However, the effective sample size in this situation is $200/[1 + (20 - 1).1] = 69.0$, which is much less than the apparent total sample size of 200! Gulliford, Ukoumunne, and Chin (1999) give an overview of estimates of the intraclass correlation to aid in the design of complex health surveys. Their data include variables on a range of lifestyle risk factors and health outcomes, for respondents clustered at the household, postal code, and health authority district levels. They report between-cluster variation at each of these levels, with intraclass correlations ranging from .0 to .3 at the household level, and being mostly smaller than .05 at the postal code level, and below .01 at the district level. Smeeth and Ng (2002) present ICCs for health-related variables for elderly patients within primary-care clinics. Their ICCs are generally small, the largest being .06 for 'difficult to keep house warm'. Smeeth and Ng (2002) list 17 other studies that report ICCs in the field of health research.

Since the design effect depends on both the intraclass correlation and the cluster size, large intraclass correlations are partly compensated by small group sizes. Conversely, small intraclass correlations at the higher levels are offset by the usually large cluster sizes at these levels. Groves (1989) also discusses the effects of cluster sampling on the standard errors in cluster samples, and concludes that the intraclass correlation is usually small, but in combination with the usual cluster sizes used in surveys they still can lead to substantial design effects.

Some of the correction procedures developed for cluster and other complex samples are quite powerful (see Skinner, Holt, & Smith, 1989). In principle such correction procedures could also be applied in analyzing multilevel data, by adjusting the standard errors of the statistical tests. However, multilevel models are multivariate models, and in general the intraclass correlation and hence the effective N is different for different variables. In addition, in most multilevel problems we do not only have clustering of individuals within groups, but we also have variables measured at all available levels, and we are interested in the relationships between all these variables. Combining variables from different levels in one statistical model is a different and more complicated problem than estimating and correcting for design effects. Multilevel models are designed to analyze variables from different levels simultaneously, using a statistical model that properly includes the various dependencies.

To provide an example of a clearly multilevel problem, consider the 'frog pond' theory that has been utilized in educational and organizational research. The 'frog pond' theory refers to the notion that a specific individual frog may be a medium sized frog in a pond otherwise filled with large frogs, or a medium sized frog in a pond otherwise filled with small frogs. Applied to education, this metaphor points out that the effect of an explanatory variable such as 'intelligence' on school career may depend on the average intelligence of the other pupils in the school. A moderately intelligent

pupil in a highly intelligent context may become demotivated and thus become an underachiever, while the same pupil in a considerably less intelligent context may gain confidence and become an overachiever. Thus, the effect of an individual pupil's intelligence depends on the average intelligence of the other pupils in the class. A popular approach in educational research to investigate 'frog pond' effects has been to aggregate variables like the pupils' IQs into group means, and then to disaggregate these group means again to the individual level. As a result, the data file contains both individual-level (global) variables and higher-level (contextual) variables in the form of disaggregated group means. Cronbach (1976; Cronbach & Webb, 1979) has suggested expressing the individual scores as deviations from their respective group means, a procedure that has become known as *centering on the group mean*, or *group mean centering*. Centering on the group mean makes very explicit that the individual scores should be interpreted relative to their group's mean. The example of the 'frog pond' theory and the corresponding practice of centering the predictor variables makes clear that combining and analyzing information from different levels within one statistical model is central to multilevel modeling.

1.3 MULTILEVEL THEORIES

Multilevel problems must be explained by multilevel theories, an area that seems underdeveloped compared to the advances made in modeling and computing machinery (see Hüttner & van den Eeden, 1993). Multilevel models in general require that the grouping criterion is clear, and that variables can be assigned unequivocally to their appropriate level. In reality, group boundaries are sometimes fuzzy and somewhat arbitrary, and the assignment of variables is not always obvious and simple. In multilevel problems, decisions about group membership and operationalizations involve a wide range of theoretical assumptions, and an equally wide range of specification problems for the auxiliary theory (Blalock, 1990; Klein & Kozlowski, 2000). If there are effects of the social context on individuals, these effects must be mediated by intervening processes that depend on characteristics of the social context. When the number of variables at the different levels is large, there are an enormous number of possible cross-level interactions. Ideally, a multilevel theory should specify which variables belong to which level, and which direct effects and cross-level interaction effects can be expected. Cross-level interaction effects between the individual and the context level require the specification of processes within individuals that cause those individuals to be differentially influenced by certain aspects of the context. Attempts to identify such processes have been made by, among others, Stinchcombe (1968), Erbring and Young (1979), and Chan (1998). The common core in these theories is that they all postulate one or more psychological processes that mediate between individual

variables and group variables. Since a global explanation by 'group telepathy' is generally not acceptable, communication processes and the internal structure of groups become important concepts. These are often measured as a 'structural variable'. In spite of their theoretical relevance, structural variables are infrequently used in multilevel research. Another theoretical area that has been largely neglected by multilevel researchers is the influence of individuals on the group. This is already visible in Durkheim's concept of sociology as a science that focuses primarily on the constraints that a society can put on its members, and disregards the influence of individuals on their society. In multilevel modeling, the focus is on models where the outcome variable is at the lowest level. Models that investigate the influence of individual variables on group outcomes are scarce. For a review of this issue see DiPrete and Forristal (1994); an example is discussed by Alba and Logan (1992). Croon and van Veldhoven (2007) discuss analysis methods for multilevel data where the outcome variable is at the highest level.

1.4 MODELS DESCRIBED IN THIS BOOK

This book treats two classes of multilevel models: multilevel regression models, and multilevel models for covariance structures.

Multilevel regression models are essentially a multilevel version of the familiar multiple regression model. As Cohen and Cohen (1983), Pedhazur (1997), and others have shown, the multiple regression model is very versatile. Using dummy coding for categorical variables, it can be used to analyze analysis of variance (ANOVA)-type of models as well as the more usual multiple regression models. Since the multilevel regression model is an extension of the classical multiple regression model, it too can be used in a wide variety of research problems.

Chapter 2 of this book contains a basic introduction to the multilevel regression model, also known as the hierarchical linear model, or the random coefficient model. Chapters 3 and 4 discuss estimation procedures, and a number of important methodological and statistical issues. They also discuss some technical issues that are not specific to multilevel regression analysis, such as centering and interpreting interactions.

Chapter 5 introduces the multilevel regression model for longitudinal data. The model is a straightforward extension of the standard multilevel regression model, but there are some specific complications, such as autocorrelated errors, which are discussed.

Chapter 6 treats the generalized linear model for dichotomous data and proportions. When the response (dependent) variable is dichotomous or a proportion, standard regression models should not be used. This chapter discusses the multilevel version of the logistic and the probit regression model.

Chapter 7 extends the generalized linear model introduced in Chapter 6 to analyze data that are ordered categorical and to data that are counts. In the context of counts, it presents models that take an overabundance of zeros into account.

Chapter 8 introduces multilevel modeling of survival or event history data. Survival models are for data where the outcome is the occurrence or nonoccurrence of a certain event, in a certain observation period. If the event has not occurred when the observation period ends, the outcome is said to be censored, since we do not know whether or not the event has taken place after the observation period ended.

Chapter 9 discusses cross-classified models. Some data are multilevel in nature, but do not have a neat hierarchical structure. Examples are longitudinal school research data, where pupils are nested within schools, but may switch to a different school in later measurements, and sociometric choice data. Multilevel models for such cross-classified data can be formulated, and estimated with standard software provided that it can handle restrictions on estimated parameters.

Chapter 10 discusses multilevel regression models for multivariate outcomes. These can also be used to estimate models that resemble confirmative factor analysis, and to assess the reliability of multilevel measurements. A different approach to multilevel confirmative factor analysis is treated in Chapter 13.

Chapter 11 describes a variant of the multilevel regression model that can be used in meta-analysis. It resembles the weighted regression model often recommended for meta-analysis. Using standard multilevel regression procedures, it is a flexible analysis tool, especially when the meta-analysis includes multivariate outcomes.

Chapter 12 deals with the sample size needed for multilevel modeling, and the problem of estimating the power of an analysis given a specific sample size. An obvious complication in multilevel power analysis is that there are different sample sizes at the distinct levels, which should be taken into account.

Chapter 13 treats some advanced methods of estimation and assessing significance. It discusses the profile likelihood method, robust standard errors for establishing confidence intervals, and multilevel bootstrap methods for estimating bias-corrected point-estimates and confidence intervals. This chapter also contains an introduction into Bayesian (MCMC) methods for estimation and inference.

Multilevel models for covariance structures, or multilevel structural equation models (SEM), are a powerful tool for the analysis of multilevel data. Recent versions of structural equation modeling software, such as EQS, LISREL, Mplus, all include at least some multilevel features. The general statistical model for multilevel covariance structure analysis is quite complicated. Chapter 14 in this book describes both a simplified statistical model proposed by Muthén (1990, 1994), and more recent developments. It explains how multilevel confirmatory factor models can be estimated with either conventional SEM software or using specialized programs. In addition, it deals with issues of calculating standardized coefficients and goodness-of-fit indices in

multilevel structural models. Chapter 15 extends this to path models. Chapter 16 describes structural models for latent curve analysis. This is a SEM approach to analyzing longitudinal data, which is very similar to the multilevel regression models treated in Chapter 5.

This book is intended as an introduction to the world of multilevel analysis. Most of the chapters on multilevel regression analysis should be readable for social scientists who have a good general knowledge of analysis of variance and classical multiple regression analysis. Some of these chapters contain material that is more difficult, but this is generally a discussion of specialized problems, which can be skipped at first reading. An example is the chapter on longitudinal models, which contains a prolonged discussion of techniques to model specific structures for the covariances between adjacent time points. This discussion is not needed to understand the essentials of multilevel analysis of longitudinal data, but it may become important when one is actually analyzing such data. The chapters on multilevel structure equation modeling obviously require a strong background in multivariate statistics and some background in structural equation modeling, equivalent to, for example, the material covered in Tabachnick and Fidell's (2007) book. Conversely, in addition to an adequate background in structural equation modeling, the chapters on multilevel structural equation modeling do not require knowledge of advanced mathematical statistics. In all these cases, I have tried to keep the discussion of the more advanced statistical techniques theoretically sound, but non-technical.

Many of the techniques and their specific software implementations discussed in this book are the subject of active statistical and methodological research. In other words: both the statistical techniques and the software tools are evolving rapidly. As a result, increasing numbers of researchers will apply increasingly advanced models to their data. Of course, researchers still need to understand the models and techniques that they use. Therefore, in addition to being an introduction to multilevel analysis, this book aims to let the reader become acquainted with some advanced modeling techniques that might be used, such as bootstrapping and Bayesian estimation methods. At the time of writing, these are specialist tools, and certainly not part of the standard analysis toolkit. But they are developing rapidly, and are likely to become more popular in applied research as well.

2

The Basic Two-Level Regression Model

The multilevel regression model has become known in the research literature under a variety of names, such as 'random coefficient model' (de Leeuw & Kreft, 1986; Longford, 1993), 'variance component model' (Longford, 1987), and 'hierarchical linear model' (Raudenbush & Bryk, 1986, 1988). Statistically oriented publications tend to refer to the model as a mixed-effects or mixed model (Littell, Milliken, Stroup, & Wolfinger, 1996). The models described in these publications are not *exactly* the same, but they are highly similar, and I will refer to them collectively as 'multilevel regression models'. They all assume that there is a hierarchical data set, with one single outcome or response variable that is measured at the lowest level, and explanatory variables at all existing levels. Conceptually, it is useful to view the multilevel regression model as a hierarchical system of regression equations. In this chapter, I will explain the multilevel regression model for two-level data, and also give an example of three-level data. Regression models with more than two levels are also used in later chapters.

2.1 EXAMPLE

Assume that we have data from J classes, with a different number of pupils n_j in each class. On the pupil level, we have the outcome variable 'popularity' (Y), measured by a self-rating scale that ranges from 0 (very unpopular) to 10 (very popular). We have two explanatory variables on the pupil level: *pupil gender* (X_1: 0 = boy, 1 = girl) and *pupil extraversion* (X_2, measured on a self-rating scale ranging from 1 to 10), and one class-level explanatory variable *teacher experience* (Z: in years, ranging from 2 to 25). There are data on 2000 pupils in 100 classes, so the average class size is 20 pupils. The data are described in Appendix A.

To analyze these data, we can set up separate regression equations in each class to predict the outcome variable Y using the explanatory variables X as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + e_{ij}. \tag{2.1}$$

Using variable labels instead of algebraic symbols, the equation reads:

$$popularity_{ij} = \beta_{0j} + \beta_{1j}\, gender_{ij} + \beta_{2j}\, extraversion_{ij} + e_{ij}. \tag{2.2}$$

In this regression equation, β_{0j} is the intercept, β_{1j} is the regression coefficient (regression slope) for the dichotomous explanatory variable gender, β_{2j} is the regression coefficient (slope) for the continuous explanatory variable extraversion, and e_{ij} is the usual residual error term. The subscript j is for the classes ($j = 1 \ldots J$) and the subscript i is for individual pupils ($i = 1 \ldots n_j$). The difference with the usual regression model is that we assume that each class has a different intercept coefficient β_{0j}, and different slope coefficients β_{1j} and β_{2j}. This is indicated in equations 2.1 and 2.2 by attaching a subscript j to the regression coefficients. The residual errors e_{ij} are assumed to have a mean of zero, and a variance to be estimated. Most multilevel software assumes that the variance of the residual errors is the same in all classes. Different authors (see Goldstein, 2003; Raudenbush & Bryk, 2002) use different systems of notation. This book uses σ_e^2 to denote the variance of the lowest level residual errors.[1]

Since the intercept and slope coefficients are random variables that vary across the classes, they are often referred to as *random* coefficients.[2] In our example, the specific values for the intercept and the slope coefficients are a class characteristic. In general, a class with a high intercept is predicted to have more popular pupils than a class with a low value for the intercept.[3] Similarly, differences in the slope coefficient for gender or extraversion indicate that the relationship between the pupils' gender or extraversion and their predicted popularity is not the same in all classes. Some classes may have a high value for the slope coefficient of gender; in these classes, the difference between boys and girls is relatively large. Other classes may have a low value for the slope coefficient of gender; in these classes, gender has a small effect on the popularity, which means that the difference between boys and girls is small. Variance in the slope for pupil extraversion is interpreted in a similar way; in classes with a large coefficient for the extraversion slope, pupil extraversion has a large impact on their popularity, and vice versa.

Across all classes, the regression coefficients β_j are assumed to have a multivariate normal distribution. The next step in the hierarchical regression model is to explain the variation of the regression coefficients β_j introducing explanatory variables at the class level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}, \tag{2.3}$$

[1] At the end of this chapter, a section explains the difference between some commonly used notation systems. Models that are more complicated sometimes need a more complicated notation system, which is introduced in the relevant chapters.

[2] Of course, we hope to explain at least some of the variation by introducing higher-level variables. Generally, we will not be able to explain all the variation, and there will be some unexplained residual variation.

[3] Since the model contains a dummy variable for gender, the precise value of the intercept reflects the predicted value for the boys (coded as zero). Varying intercepts shift the average value for the entire class, both boys and girls.

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21} Z_j + u_{2j}. \tag{2.4}$$

Equation 2.3 predicts the average popularity in a class (the intercept β_{0j}) by the teacher's experience (Z). Thus, if γ_{01} is positive, the average popularity is higher in classes with a more experienced teacher. Conversely, if γ_{01} is negative, the average popularity is lower in classes with a more experienced teacher. The interpretation of the equations under 2.4 is a bit more complicated. The first equation under 2.4 states that the *relationship*, as expressed by the slope coefficient β_{1j}, between the popularity (Y) and the gender (X) of the pupil depends on the amount of experience of the teacher (Z). If γ_{11} is positive, the gender effect on popularity is larger with experienced teachers. Conversely, if γ_{11} is negative, the gender effect on popularity is smaller with experienced teachers. Similarly, the second equation under 2.4 states, if γ_{21} is positive, that the effect of extraversion is larger in classes with an experienced teacher. Thus, the amount of experience of the teacher acts as a *moderator variable* for the relationship between popularity and gender or extraversion; this relationship varies according to the value of the moderator variable.

The u-terms u_{0j}, u_{1j} and u_{2j} in equations 2.3 and 2.4 are (random) residual error terms at the class level. These residual errors u_j are assumed to have a mean of zero, and to be independent from the residual errors e_{ij} at the individual (pupil) level. The variance of the residual errors u_{0j} is specified as $\sigma^2_{u_0}$, and the variance of the residual errors u_{1j} and u_{2j} is specified as $\sigma^2_{u_1}$ and $\sigma^2_{u_2}$. The *covariances* between the residual error terms are denoted by $\sigma_{u_{01}}$, $\sigma_{u_{02}}$ and $\sigma_{u_{12}}$, which are generally *not* assumed to be zero.

Note that in equations 2.3 and 2.4 the regression coefficients γ are not assumed to vary across classes. They therefore have no subscript j to indicate to which class they belong. Because they apply to *all* classes, they are referred to as *fixed* coefficients. All between-class variation left in the β coefficients, after predicting these with the class variable Z_j, is assumed to be residual error variation. This is captured by the residual error terms u_j, which do have subscripts j to indicate to which class they belong.

Our model with two pupil-level and one class-level explanatory variable can be written as a single complex regression equation by substituting equations 2.3 and 2.4 into equation 2.1. Rearranging terms gives:

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{10} X_{1ij} + \gamma_{20} X_{2ij} + \gamma_{01} Z_j + \gamma_{11} X_{1ij} Z_j + \gamma_{21} X_{2ij} Z_j \\ & + u_{1j} X_{1ij} + u_{2j} X_{2ij} + u_{0j} + e_{ij}. \end{aligned} \tag{2.5}$$

Using variable labels instead of algebraic symbols, we have:

$$popularity_{ij} = \gamma_{00} + \gamma_{10}\, gender_{ij} + \gamma_{20}\, extraversion_{ij} + \gamma_{01}\, experience_j + \gamma_{11}\, gender_{ij} \times experience_j + \gamma_{21}\, extraversion_{ij} \times experience_j + u_{1j}\, gender_{ij} + u_{2j}\, extraversion_{ij} + u_{0j} + e_{ij}.$$

The segment $[\gamma_{00} + \gamma_{10} X_{1ij} + \gamma_{20} X_{2ij} + \gamma_{01} Z_j + \gamma_{11} X_{1ij} Z_j + \gamma_{11} X_{2ij} Z_j]$ in equation 2.5 contains the fixed coefficients. It is often called the fixed (or deterministic) part of the model. The segment $[u_{1j} X_{1ij} + u_{2j} X_{2ij} + u_{0j} + e_{ij}]$ in equation 2.5 contains the random error terms, and it is often called the random (or stochastic) part of the model. The terms $X_{1i} Z_{jj}$ and $X_{2ij} Z_j$ are interaction terms that appear in the model as a consequence of modeling the varying regression slope β_j of a pupil-level variable X_{ij} with the class-level variable Z_j. Thus, the moderator effect of Z on the relationship between the dependent variable Y and the predictor X is expressed in the single equation version of the model as a *cross-level interaction*. The interpretation of interaction terms in multiple regression analysis is complex, and this is treated in more detail in Chapter 4. In brief, the point made in Chapter 4 is that the substantive interpretation of the coefficients in models with interactions is much simpler if the variables making up the interaction are expressed as deviations from their respective means.

Note that the random error terms u_{1j} are connected to X_{ij}. Since the explanatory variable X_{ij} and the corresponding error term u_j are multiplied, the resulting total error will be different for different values of the explanatory variable X_{ij}, a situation that in ordinary multiple regression analysis is called 'heteroscedasticity'. The usual multiple regression model assumes 'homoscedasticity', which means that the variance of the residual errors is independent of the values of the explanatory variables. If this assumption is not true, ordinary multiple regression does not work very well. This is another reason why analyzing multilevel data with ordinary multiple regression techniques does not work well.

As explained in the introduction in Chapter 1, multilevel models are needed because with grouped data observations from the same group are generally more similar to each other than the observations from different groups, and this violates the assumption of independence of all observations. The amount of dependence can be expressed as a correlation coefficient: the intraclass correlation. The methodological literature contains a number of different formulas to estimate the intraclass correlation ρ. For example, if we use one-way analysis of variance with the grouping variable as independent variable to test the group effect on our outcome variable, the intraclass correlation is given by $\rho = [MS(B) - MS(error)]/[MS(B)+(n-1) \times MS(error)]$, where $MS(B)$ is the between groups mean square and n is the common group size. Shrout and Fleiss (1979) give an overview of formulas for the intraclass correlation for a variety of research designs.

If we have simple hierarchical data, the multilevel regression model can also be used to produce an estimate of the intraclass correlation. The model used for this

purpose is a model that contains no explanatory variables at all, the so-called *intercept-only* model. The intercept-only model is derived from equations 2.1 and 2.3 as follows. If there are no explanatory variables X at the lowest level, equation 2.1 reduces to:

$$Y_{ij} = \beta_{0j} + e_{ij}. \tag{2.6}$$

Likewise, if there are no explanatory variables Z at the highest level, equation 2.3 reduces to:

$$\beta_{0j} = \gamma_{00} + u_{0j}. \tag{2.7}$$

We find the single equation model by substituting 2.7 into 2.6:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}. \tag{2.8}$$

We could also have found equation 2.8 by removing all terms that contain an X or Z variable from equation 2.5. The intercept-only model of equation 2.8 does not explain any variance in Y. It only decomposes the variance into two independent components: σ_e^2, which is the variance of the lowest-level errors e_{ij}, and σ_{u0}^2, which is the variance of the highest-level errors u_{0j}. Using this model, we can define the intraclass correlation ρ by the equation:

$$\rho = \frac{\sigma_{u_0}^2}{\sigma_{u_0}^2 + \sigma_e^2}. \tag{2.9}$$

The intraclass correlation ρ indicates the proportion of the variance explained by the grouping structure in the population. Equation 2.9 simply states that the intraclass correlation is the proportion of group-level variance compared to the total variance.[4] The intraclass correlation ρ can also be interpreted as the expected correlation between two randomly drawn units that are in the same group.

Ordinary multiple regression analysis uses an estimation technique called ordinary least squares, abbreviated as OLS. The statistical theory behind the multilevel regression model is more complex, however. Based on observed data, we want to estimate the parameters of the multilevel regression model: the regression coefficients and the variance components. The usual estimators in multilevel regression analysis are

[4] The intraclass correlation is an estimate of the proportion of group-level variance in the *population*. The proportion of group-level variance in the *sample* is given by the correlation ratio η^2 (eta-squared, see Tabachnick & Fidell, 2007, p. 54): $\eta^2 = SS(B)/SS(Total)$.

maximum likelihood (ML) estimators. Maximum likelihood estimators estimate the parameters of a model by providing estimated values for the population parameters that maximize the so-called 'likelihood function': the function that describes the probability of observing the sample data, given the specific values of the parameter estimates. Simply put, ML estimates are those parameter estimates that maximize the probability of finding the sample data that we have actually found. For an accessible introduction to maximum likelihood methods see Eliason (1993).

Maximum likelihood estimation includes procedures to generate standard errors for most of the parameter estimates. These can be used in significance testing, by computing the test statistic Z: $Z = parameter / (st.error\ param.)$. This statistic is referred to the standard normal distribution, to establish a p-value for the null-hypothesis that the population value of that parameter is zero. The maximum likelihood procedure also produces a statistic called the *deviance*, which indicates how well the model fits the data. In general, models with a lower deviance fit better than models with a higher deviance. If two models are *nested*, meaning that a specific model can be derived from a more general model by removing parameters from that general model, the deviances of the two models can be used to compare their fit statistically. For nested models, the difference in deviance has a chi-square distribution with degrees of freedom equal to the difference in the number of parameters that are estimated in the two models. The deviance test can be used to perform a formal chi-square test, in order to test whether the more general model fits significantly better than the simpler model. The chi-square test of the deviances can also be used to good effect to explore the importance of a set of random effects, by comparing a model that contains these effects against a model that excludes them.

2.2 AN EXTENDED EXAMPLE

The intercept-only model is useful as a null model that serves as a benchmark with which other models are compared. For our pupil popularity example data, the intercept-only model is written as:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}.$$

The model that includes pupil gender, pupil extraversion and teacher experience, but not the cross-level interactions, is written as:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{1ij} + \gamma_{20} X_{2ij} + \gamma_{01} Z_j + u_{1j} X_{1ij} + u_{2j} X_{2ij} + u_{0j} + e_{ij},$$

or, using variable names instead of algebraic symbols:

$$popularity_{ij} = \gamma_{00} + \gamma_{10}\, gender_{ij} + \gamma_{20}\, extraversion_{ij} + \gamma_{01}\, experience_j + u_{1j}\, gender_{ij} + u_{2j}\, extraversion_{ij} + u_{0j} + e_{ij}.$$

Table 2.1 Intercept-only model and model with explanatory variables

Model	M0: intercept only	M1: with predictors
Fixed part	Coefficient (s.e.)	Coefficient (s.e.)
Intercept	5.08 (.09)	0.74 (.20)
Pupil gender		1.25 (.04)
Pupil extraversion		0.45 (.03)
Teacher experience		0.09 (.01)
Random part[a]		
σ^2_e	1.22 (.04)	0.55 (.02)
σ^2_{u0}	0.69 (.11)	1.28 (.47)
σ^2_{u1}		0.00 (–)
σ^2_{u2}		0.03 (.008)
Deviance	6327.5	4812.8

[a] For simplicity the covariances are not included.

Table 2.1 presents the parameter estimates and standard errors for both models.[5] In this table, the intercept-only model estimates the intercept as 5.08, which is simply the average popularity across all classes and pupils. The variance of the pupil-level residual errors, symbolized by σ^2_e, is estimated as 1.22. The variance of the class-level residual errors, symbolized by σ^2_{u0}, is estimated as 0.69. All parameter estimates are much larger than the corresponding standard errors, and calculation of the Z-test shows that they are all significant at $p < .005$.[6] The intraclass correlation, calculated by equation 2.9 as $\rho = \sigma^2_{u0} \big/ \left(\sigma^2_{u0} + \sigma^2_e\right)$, is 0.69/1.91, which equals .36. Thus, 36% of the variance of the popularity scores is at the group level, which is very high. Since the intercept-only model contains no explanatory variables, the residual variances represent unexplained error variance. The deviance reported in Table 2.1 is a measure of

[5] For reasons to be explained later, different options for the details of the maximum likelihood procedure may result in slightly different estimates. So, if you re-analyze the example data from this book, the results may differ slightly from the results given here. However, these differences should never be so large that you would draw entirely different conclusions.

[6] Testing variances is preferably done with a test based on the deviance, which is explained in Chapter 3.

model misfit; when we add explanatory variables to the model, the deviance is expected to go down.

The second model in Table 2.1 includes pupil gender and extraversion and teacher experience as explanatory variables. The regression coefficients for all three variables are significant. The regression coefficient for pupil gender is 1.25. Since pupil gender is coded 0 = boy, 1 = girl, this means that on average the girls score 1.25 points higher on the popularity measure. The regression coefficient for pupil extraversion is 0.45, which means that with each scale point higher on the extraversion measure, the popularity is expected to increase by 0.45 scale points. The regression coefficient for teacher experience is 0.09, which means that for each year of experience of the teacher, the average popularity score of the class goes up by 0.09 points. This does not seem very much, but the teacher experience in our example data ranges from 2 to 25 years, so the predicted difference between the least experienced and the most experienced teacher is $(25 - 2 =) 23 \times 0.09 = 2.07$ points on the popularity measure. We can use the standard errors of the regression coefficients reported in Table 2.1 to construct a 95% confidence interval. For the regression coefficient of pupil gender, the 95% confidence interval runs from 1.17 to 1.33, the confidence interval for pupil extraversion runs from 0.39 to 0.51, and the 95% confidence interval for the regression coefficient of teacher experience runs from 0.07 to 0.11.

The model with the explanatory variables includes variance components for the regression coefficients of pupil gender and pupil extraversion, symbolized by σ^2_{u1} and σ^2_{u2} in Table 2.1. The variance of the regression coefficients for pupil extraversion across classes is estimated as 0.03, with a standard error of .008. The simple Z-test (Z = 3.75) results in a (one sided) p-value of $p < .001$, which is clearly significant. The variance of the regression coefficients for pupil gender is estimated as 0.00. This variance component is clearly not significant, so the hypothesis that the regression slopes for pupil gender vary across classes is not supported by the data. Therefore we can remove the residual variance term for the gender slopes from the model.[7] Table 2.2 presents the estimates for the model with a fixed slope for the effect of pupil gender. Table 2.2 also includes the covariance between the class-level errors for the intercept and the extraversion slope. These covariances are rarely interpreted, and for that reason often not included in the reported tables. However, as Table 2.2 demonstrates, they can be quite large and significant, so as a rule they are always included in the model.

The significant variance of the regression slopes for pupil extraversion implies that we should not interpret the estimated value of 0.45 without considering this

[7] Multilevel software deals with the problem of zero variances in different ways. Most software inserts a zero which may or may not be flagged as a redundant parameter. In general, such zero variances should be removed from the model, and the resulting new model must be re-estimated.

Table 2.2 Model with explanatory variables, extraversion slope random

Model	M1: with predictors
Fixed part	Coefficient (s.e.)
Intercept	0.74 (.20)
Pupil gender	1.25 (.04)
Pupil extraversion	0.45 (.02)
Teacher experience	0.09 (.01)
Random part	
σ_e^2	0.55 (.02)
σ_{u0}^2	1.28 (.28)
σ_{u2}^2	0.03 (.008)
$\sigma_{u_{02}}$	–0.18 (.05)
Deviance	4812.8

variation. In an ordinary regression model, without multilevel structure, the value of 0.45 means that for each point of difference on the extraversion scale, pupil popularity goes up by 0.45, for all pupils in all classes. In our multilevel model, the regression coefficient for pupil gender varies across the classes, and the value of 0.45 is just the expected value (the mean) across all classes. The varying regression slopes for pupil extraversion are assumed to follow a normal distribution. The variance of this distribution is in our example estimated as 0.034. Interpretation of this variation is easier when we consider the standard deviation, which is the square root of the variance or 0.18 in our example data. A useful characteristic of the standard deviation is that with normally distributed observations about 67% of the observations lie between one standard deviation below and one above the mean, and about 95% of the observations lie between two standard deviations below and above the mean. If we apply this to the regression coefficients for pupil gender, we conclude that about 67% of the regression coefficients are expected to lie between (0.45 – 0.18 =) 0.27 and (0.45 + 0.18 =) 0.63, and about 95% are expected to lie between (0.45 – 0.37 =) 0.08 and (0.45 + 0.37 =) 0.82. The more precise value of $Z_{.975} = 1.96$ leads to the 95% predictive interval calculated as 0.09 to 0.81. We can also use the standard normal distribution to estimate the percentage of regression coefficients that are negative. As it turns out, if the mean regression coefficient for pupil extraversion is 0.45, given the estimated slope variance, less than 1% of the classes are expected to have a regression coefficient that is actually negative. Note that the 95% interval computed here is totally different from the 95% confidence

interval for the regression coefficient of pupil extraversion, which runs from 0.41 to 0.50. The 95% confidence interval applies to γ_{20}, the mean value of the regression coefficients across all the classes. The 95% interval calculated here is the 95% *predictive interval*, which expresses that 95% of the regression coefficients of the variable 'pupil extraversion' in the classes are predicted to lie between 0.09 and 0.81.

Given the significant variance of the regression coefficient of pupil extraversion across the classes it is attractive to attempt to predict its variation using class-level variables. We have one class-level variable: teacher experience. The individual-level regression equation for this example, using variable labels instead of symbols, is given by equation 2.10:

$$popularity_{ij} = \beta_{0j} + \beta_1\, gender_{ij} + \beta_{2j}\, extraversion_{ij} + e_{ij}. \tag{2.10}$$

The regression coefficient β_1 for pupil gender does not have a subscript j, because it is not assumed to vary across classes. The regression equations predicting β_{0j}, the intercept in class j, and β_{2j}, the regression slope of pupil extraversion in class j, are given by equation 2.3 and 2.4, which are rewritten below using variable labels:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} t.exp_j + u_{0j} \\ \beta_{2j} &= \gamma_{20} + \gamma_{21} t.exp_j + u_{2j}. \end{aligned} \tag{2.11}$$

By substituting 2.11 into 2.10 we get:

$$\begin{aligned} popularity_{ij} = {} & \gamma_{00} + \gamma_{10}\, gender_{ij} + \gamma_{20}\, extraversion_{ij} + \gamma_{01} t.exp_j \\ & + \gamma_{21} extraversion_{ij}\, t.exp_j + u_{2j}\, extraversion_{ij} + u_{0j} + e_{ij} \end{aligned} \tag{2.12}$$

The algebraic manipulations of the equations above make clear that to explain the variance of the regression slopes β_{2j}, we need to introduce an interaction term in the model. This interaction, between the variables pupil extraversion and teacher experience, is a cross-level interaction, because it involves explanatory variables from different levels. Table 2.3 presents the estimates from a model with this cross-level interaction. For comparison, the estimates for the model without this interaction are also included in Table 2.3.

The estimates for the fixed coefficients in Table 2.3 are similar for the effect of pupil gender, but the regression slopes for pupil extraversion and teacher experience are considerably larger in the cross-level model. The interpretation remains the same: extraverted pupils are more popular. The regression coefficient for the cross-level interaction is –0.03, which is small but significant. This interaction is formed by multiplying the scores for the variables 'pupil extraversion' and 'teacher experience', and the negative value means that with experienced teachers, the advantage of being

Table 2.3 Model without and with cross-level interaction

Model	M1A: main effects	M2: with interaction
Fixed part	Coefficient (s.e.)	Coefficient (s.e.)
Intercept	0.74 (.20)	–1.21 (.27)
Pupil gender	1.25 (.04)	1.24 (.04)
Pupil extraversion	0.45 (.02)	0.80 (.04)
Teacher experience	0.09 (.01)	0.23 (.02)
Extra × T.exp		–0.03 (.003)
Random part		
σ^2_e	0.55 (.02)	0.55 (.02)
σ^2_{u0}	1.28 (.28)	0.45 (.16)
σ^2_{u2}	0.03 (.008)	0.005 (.004)
$\sigma_{u_{02}}$	–0.18 (.05)	–0.03 (.02)
Deviance	4812.8	4747.6

extraverted is smaller than expected from the direct effects only. Thus, the difference between extraverted and introverted pupils is smaller with more experienced teachers.

Comparison of the other results between the two models shows that the variance component for pupil extraversion goes down from 0.03 in the direct effects model to 0.005 in the cross-level model. Apparently, the cross-level model explains some of the variation of the slopes for pupil extraversion. The deviance also goes down, which indicates that the model fits better than the previous model. The other differences in the random part are more difficult to interpret. Much of the difficulty in reconciling the estimates in the two models in Table 2.3 stems from adding an interaction effect between variables that have not been centered. This issue is discussed in more detail in Chapter 4.

The coefficients in the tables are all unstandardized regression coefficients. To interpret them properly, we must take the scale of the explanatory variables into account. In multiple regression analysis, and structural equation models, for that matter, the regression coefficients are often standardized because that facilitates the interpretation when one wants to compare the effects of different variables within one sample. Only if the goal of the analysis is to compare parameter estimates from different samples to each other, should one always use unstandardized coefficients. To standardize the regression coefficients, as presented in Table 2.1 or Table 2.3, one could standardize all variables before putting them into the multilevel analysis. However, this would in general also change the estimates of the variance components. This may not

be a bad thing in itself, because standardized variables are also centered on their overall mean. Centering explanatory variables has some distinct advantages, which are discussed in Chapter 4. Even so, it is also possible to derive the standardized regression coefficients from the unstandardized coefficients:

$$standardized\ coefficient = \frac{unstandardized\ coefficient \times stand.dev.explanatory\ var.}{stand.dev.outcome\ var.} \quad (2.13)$$

In our example data, the standard deviations are: 1.38 for popularity, 0.51 for gender, 1.26 for extraversion, and 6.55 for teacher experience. Table 2.4 presents the unstandardized and standardized coefficients for the second model in Table 2.2. It also presents the estimates that we obtain if we first standardize all variables, and then carry out the analysis.

Table 2.4 Comparing unstandardized and standardized estimates

Model	Standardization using 2.13		Standardized variables
Fixed part	Coefficient (s.e.)	Standardized	Coefficient (s.e.)
Intercept	0.74 (.20)	–	–0.03 (.04)
Pupil gender	1.25 (.04)	0.46	0.45 (.01)
Pupil extraversion	0.45 (.02)	0.41	0.41 (.02)
Teacher experience	0.09 (.01)	0.43	0.43 (.04)
Random part			
σ_e^2	0.55 (.02)		0.28 (.01)
σ_{u0}^2	1.28 (.28)		0.15 (.02)
σ_{u2}^2	0.03 (.008)		0.03 (.01)
$\sigma_{u_{02}}$	–0.18 (.05)		–0.01 (.01)
Deviance	4812.8		3517.2

Table 2.4 shows that the standardized regression coefficients are almost the same as the coefficients estimated for standardized variables. The small differences in Table 2.4 are simply a result of rounding errors. However, if we use standardized variables in our analysis, we find very different variance components and a very different value for the deviance. This is not only the effect of scaling the variables differently, which becomes clear if we realize that the covariance between the slope for pupil extraversion and the intercept is significant for the unstandardized variables, but not significant for

the standardized variables. This kind of difference in results is general. The fixed part of the multilevel regression model is invariant for linear transformations, just like the regression coefficients in the ordinary single-level regression model. This means that if we change the scale of our explanatory variables, the regression coefficients and the corresponding standard errors change by the same multiplication factor, and all associated *p*-values remain exactly the same. However, the random part of the multilevel regression model is not invariant for linear transformations. The estimates of the variance components in the random part can and do change, sometimes dramatically. This is discussed in more detail in section 4.2 in Chapter 4. The conclusion to be drawn here is that, if we have a complicated random part, including random components for regression slopes, we should think carefully about the scale of our explanatory variables. If our only goal is to present standardized coefficients in addition to the unstandardized coefficients, applying equation 2.13 is safer than transforming our variables. On the other hand, we may estimate the unstandardized results, including the random part and the deviance, and then re-analyze the data using standardized variables, merely using this analysis as a computational trick to obtain the standardized regression coefficients without having to do hand calculations.

2.3 INSPECTING RESIDUALS

Inspection of residuals is a standard tool in multiple regression analysis to examine whether assumptions of normality and linearity are met (see Stevens, 2009; Tabachnick & Fidell, 2007). Multilevel regression analysis also assumes normality and linearity. Since the multilevel regression model is more complicated than the ordinary regression model, checking such assumptions is even more important. For example, Bauer and Cai (2009) show that neglecting a nonlinear relationship may result in spuriously high estimates of slope variances and cross-level interaction effects. Inspection of the residuals is one way to investigate linearity and homoscedasticity. There is one important difference from ordinary regression analysis; we have more than one residual, in fact, we have residuals for each random effect in the model. Consequently, many different residuals plots can be made.

2.3.1 Examples of residuals plots

The equation below represents the one-equation version of the direct effects model for our example data. This is the multilevel model without the cross-level interaction. Since the interaction explains part of the extraversion slope variance, a model that does not include this interaction produces a graph that displays the actual slope variation more fully.

$$popularity_{ij} = \gamma_{00} + \gamma_{10}\, gender_{ij} + \gamma_{20}\, extraversion_{ij} + \gamma_{01}\, experience_j + u_{2j}\, extraversion_{ij} + u_{0j} + e_{ij}$$

In this model, we have three residual error terms: e_{ij}, u_{0j}, and u_{2j}. The e_{ij} are the residual prediction errors at the lowest level, similar to the prediction errors in ordinary single-level multiple regression. A simple boxplot of these residuals will enable us to identify extreme outliers. An assumption that is usually made in multilevel regression analysis is that the variance of the residual errors is the same in all groups. This can be assessed by computing a one-way analysis of variance of the groups on the absolute values of the residuals, which is the equivalent of Levene's test for equality of variances in analysis of variance (Stevens, 2009). Raudenbush and Bryk (2002) describe a chi-square test that can be used for the same purpose.

The u_{0j} are the residual prediction errors at the group level, which can be used in ways analogous to the investigation of the lowest-level residuals e_{ij}. The u_{2j} are the residuals of the regression slopes across the groups. By plotting the regression slopes for the various groups, we get a visual impression of how much the regression slopes actually differ, and we may also be able to identify groups which have a regression slope that is wildly different from the others.

To test the normality assumption, we can plot standardized residuals against their normal scores. If the residuals have a normal distribution, the plot should show a straight diagonal line. Figure 2.1 is a scatterplot of the standardized level 1 residuals, calculated for the final model including cross-level interaction, against their normal scores. The graph indicates close conformity to normality, and no extreme outliers. Similar plots can be made for the level 2 residuals.

We obtain a different plot, if we plot the residuals against the predicted values of the outcome variable popularity, using the fixed part of the multilevel regression model for the prediction. Such a scatter plot of the residuals against the predicted values provides information about possible failure of normality, nonlinearity, and heteroscedasticity. If these assumptions are met, the plotted points should be evenly divided above and below their mean value of zero, with no strong structure (see Tabachnick & Fidell, 2007, p. 162). Figure 2.2 shows this scatter plot for the level 1 residuals. For our example data, the scatter plot in Figure 2.2 does not indicate strong violations of the assumptions.

Similar scatter plots can be made for the second-level residuals for the intercept and the slope of the explanatory variable pupil extraversion. As an illustration, Figure 2.3 shows the scatterplots of the level 2 residuals around the average intercept and around the average slope of pupil extraversion against the predicted values of the outcome variable popularity. Both scatterplots indicate that the assumptions are reasonably met.

An interesting plot that can be made using the level 2 residuals is a plot of the

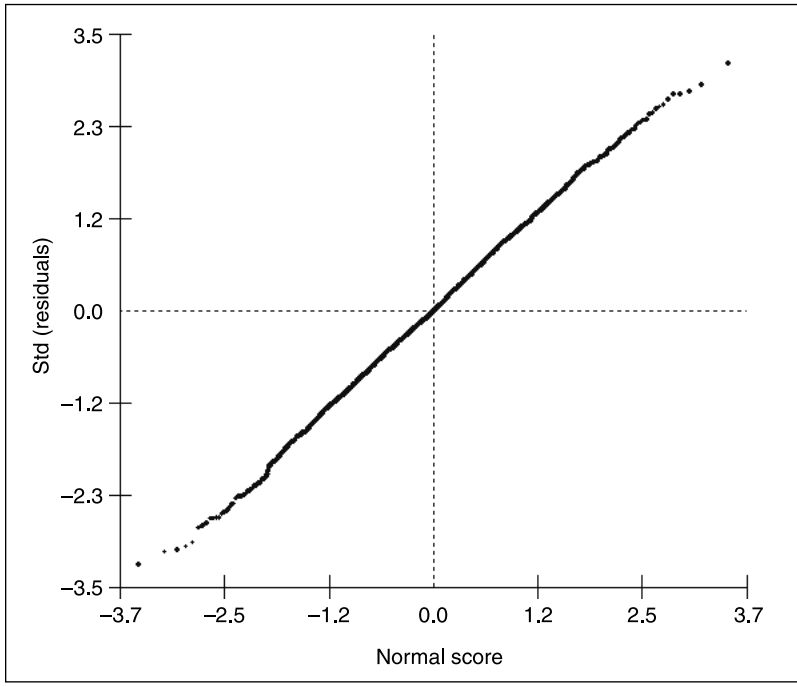

Figure 2.1 Plot of level 1 standardized residuals against normal scores.

residuals against their rank order, with an added error bar. In Figure 2.4, an error bar frames each point estimate, and the classes are sorted in rank order of the residuals. The error bars represent the confidence interval around each estimate, constructed by multiplying its standard error by 1.39 instead of the more usual 1.96. Using 1.39 as the multiplication factor results in confidence intervals with the property that if the error bars of two classes do not overlap, they have significantly different residuals at the 5% level (Goldstein, 2003). For a discussion of the construction and use of these error bars see Goldstein and Healy (1995) and Goldstein and Spiegelhalter (1996). In our example, this plot, sometimes called the *caterpillar* plot, shows some outliers at each end. This gives an indication of exceptional residuals for the intercept. A logical next step would be to identify the classes at the extremes of the rank order, and to seek a post hoc interpretation of what makes these classes different.

Examining residuals in multivariate models presents us with a problem. For instance, the residuals should show a nice normal distribution, which implies an absence of extreme outliers. However, this applies to the residuals after including all important explanatory variables and relevant parameters in the model. If we analyze a

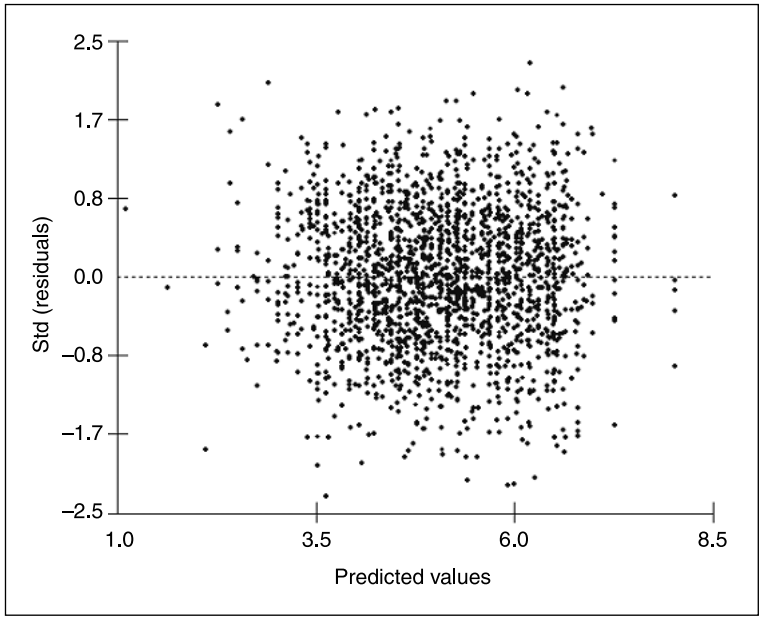

Figure 2.2 Level 1 standardized residuals plotted against predicted popularity.

sequence of models, we have a series of different residuals for each model, and scrutinizing them all at each step is not always practical. On the other hand, our decision to include a specific variable or parameter in our model might well be influenced by a violation of some assumption. Although there is no perfect solution to this dilemma, a reasonable approach is to examine the two residual terms in the intercept-only model, to find out if there are gross violations of the assumptions of the model. If there are, we should accommodate them, for instance by applying a normalizing transformation, by deleting certain individuals or groups from our data set, or by including a dummy variable that indicates a specific outlying individual or group. When we have determined our final model, we should make a more thorough examination of the various residuals. If we detect gross violations of assumptions, these should again be accommodated, and the model should be estimated again. Of course, after accommodating an extreme outlier, we might find that a previously significant effect has disappeared, and that we need to change our model again. Procedures for model exploration and detection of violations in ordinary multiple regression are discussed, for instance, in Tabachnick and Fidell (2007) or Field (2009). In multilevel regression, the same procedures apply, but the analyses are more complicated because we have to examine more than one set of residuals, and must distinguish between multiple levels.

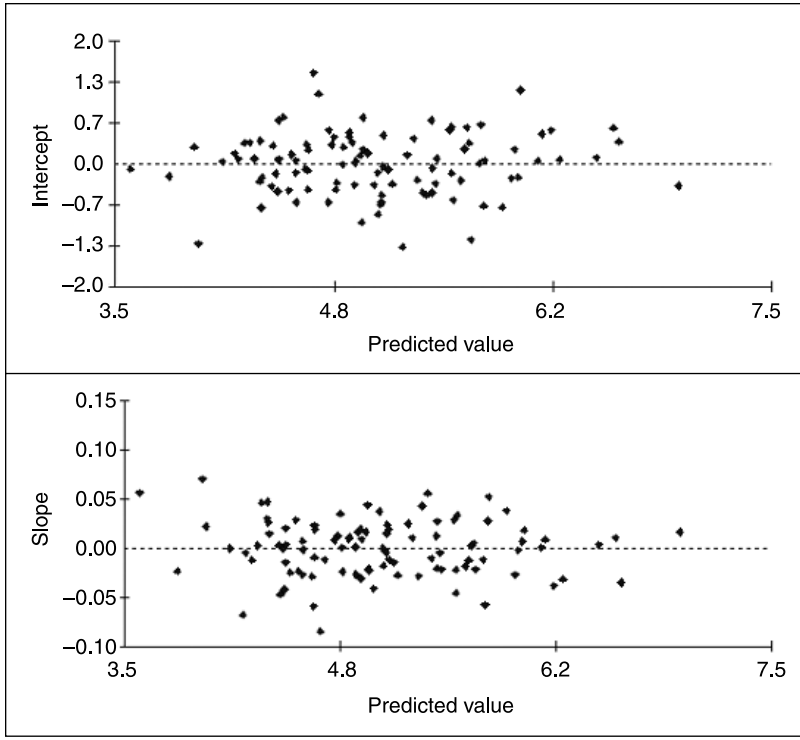

Figure 2.3 Level 2 residuals plotted against predicted popularity.

As mentioned at the beginning of this section, graphs can be useful in detecting outliers and nonlinear relations. However, an observation may have an undue effect on the outcome of a regression analysis without being an obvious outlier. Figure 2.5, a scatter plot of the so-called Anscombe data (Anscombe, 1973), illustrates this point. There is one data point in Figure 2.5, which by itself almost totally determines the regression line. Without this one observation, the regression line would be very different. Yet, when the residuals are inspected, it does not show up as an obvious outlier.

In ordinary regression analysis, various measures have been proposed to indicate the influence of individual observations on the outcome (see Tabachnick & Fidell, 2007). In general, such *influence* or *leverage* measures are based on a comparison of the estimates when a specific observation is included in the data or not. Langford and Lewis (1998) discuss extensions of these influence measures for the multilevel regression model. Since most of these measures are based on comparison of estimates with and without a specific observation, it is difficult to calculate them by hand. However, if

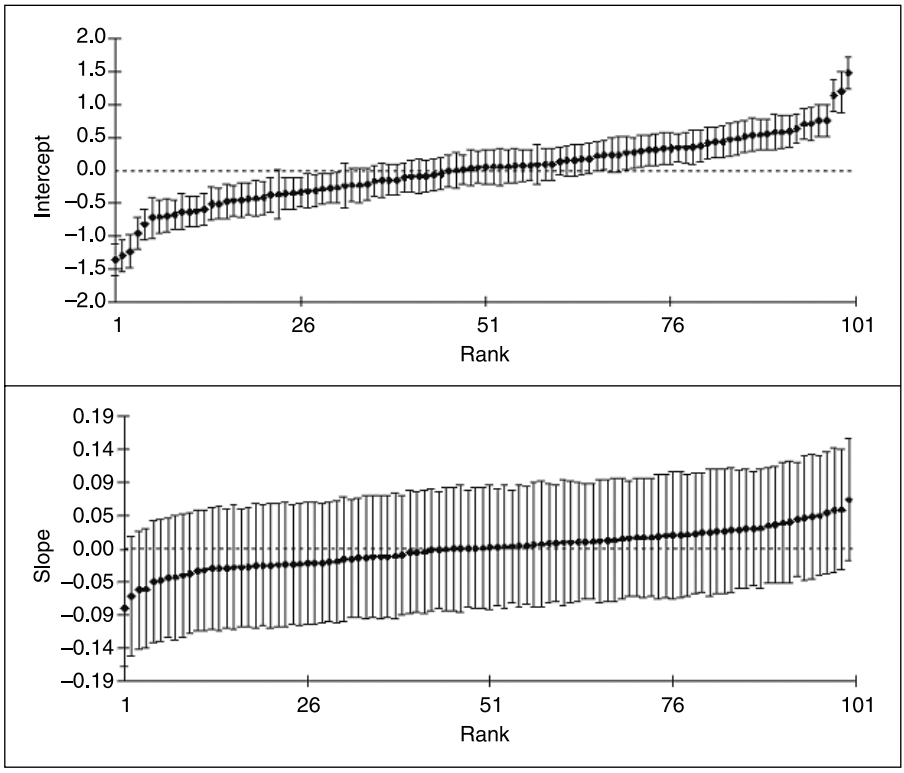

Figure 2.4 Error bar plot of level 2 residuals.

the software offers the option to calculate influence measures, it is advisable to do so. If a unit (individual or group) has a large value for the influence measure, that specific unit has a large influence on the values of the regression coefficients. It is useful to inspect cases with extreme influence values for possible violations of assumptions, or even data errors.

2.3.2 Examining slope variation: OLS and shrinkage estimators

The residuals can be added to the average values of the intercept and slope, to produce predictions of the intercepts and slopes in different groups. These can also be plotted.

For example, Figure 2.6 plots the 100 regression slopes for the explanatory variable pupil extraversion in the 100 classes. It is clear that for most classes the effect is strongly positive: extravert pupils tend to be more popular in all classes. It is also clear that in some classes the relationship is more pronounced than in other classes. Most of

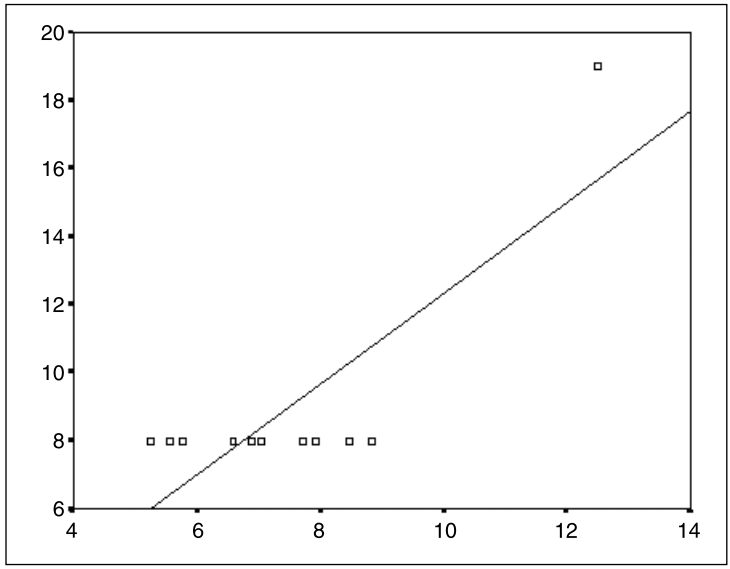

Figure 2.5 Regression line determined by one single observation.

the regression slopes are not very different from the others, although there are a few slopes that are clearly different. It could be useful to examine the data for these classes in more detail, to find out if there is a reason for the unusual slopes.

The predicted intercepts and slopes for the 100 classes are not identical to the values we would obtain if we carried out 100 separate ordinary regression analyses in each of the 100 classes, using standard ordinary least squares (OLS) techniques. If we were to compare the results from 100 separate OLS regression analyses to the values obtained from a multilevel regression analysis, we would find that the results from the separate analyses are more variable. This is because the multilevel estimates of the regression coefficients of the 100 classes are weighted. They are so-called Empirical Bayes (EB) or *shrinkage* estimates: a weighted average of the specific OLS estimate in each class and the overall regression coefficient, estimated for all similar classes.

As a result, the regression coefficients are *shrunk* back towards the mean coefficient for the whole data set. The shrinkage weight depends on the reliability of the estimated coefficient. Coefficients that are estimated with small accuracy shrink more than very accurately estimated coefficients. Accuracy of estimation depends on two factors: the group sample size, and the distance between the group-based estimate and the overall estimate. Estimates for small groups are less reliable, and shrink more than estimates for large groups. Other things being equal, estimates that are very far from the overall estimate are assumed less reliable, and they shrink more than estimates that

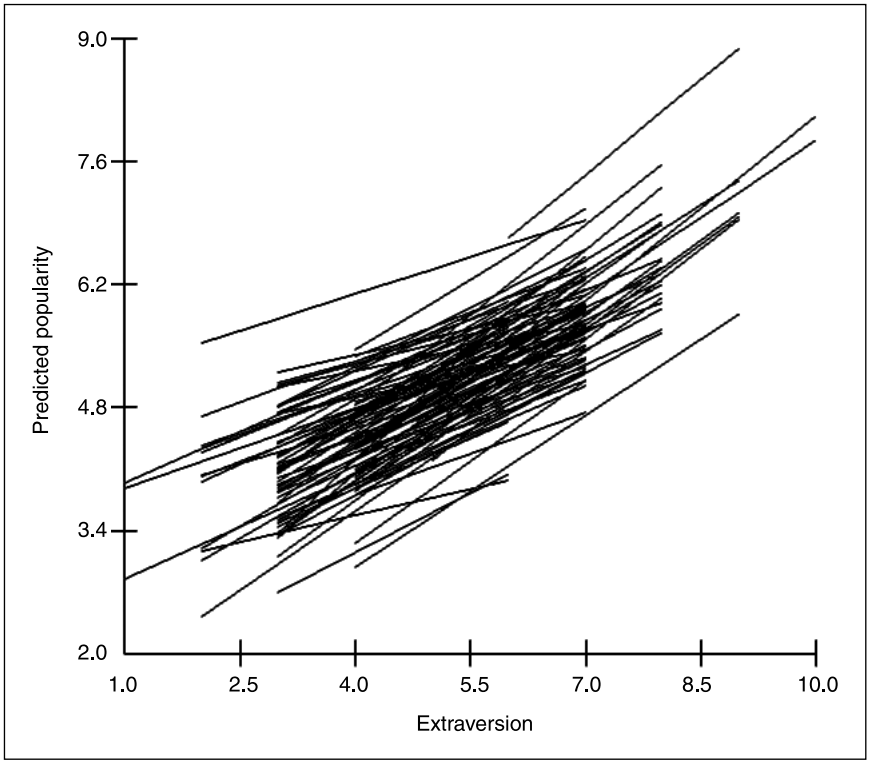

Figure 2.6 Plot of the 100 class regression slopes for pupil extraversion.

are close to the overall average. The statistical method used is called *empirical Bayes estimation*. Because of this shrinkage effect, empirical Bayes estimators are biased. However, they are usually more precise, a property that is often more useful than being unbiased (see Kendall, 1959).

The equation to form the empirical Bayes estimate of the intercepts is given by:

$$\hat{\beta}_{0j}^{\mathrm{EB}} = \lambda_j \hat{\beta}_{0j}^{\mathrm{OLS}} + (1 - \lambda_j)\gamma_{00}, \tag{2.14}$$

where λ_j is the reliability of the OLS estimate $\beta_{0j}^{\mathrm{OLS}}$ as an estimate of β_{0j}, which is given by the equation $\lambda_j = \sigma_{u_0}^2 \Big/ \left(\sigma_{u_0}^2 + \sigma_e^2 \Big/ n_j\right)$ (Raudenbush & Bryk, 2002), and γ_{00} is the overall intercept. The reliability λ_j is close to 1.0 when the group sizes are large and/or the variability of the intercepts across groups is large. In these cases, the overall estimate γ_{00} is not a good indicator of each group's intercept. If the group sizes are small and there is little variation across groups, the reliability λ_j is close to 0.0, and more

weight is put on the overall estimate γ_{00}. Equation 2.14 makes clear that, since the OLS estimates are unbiased, the empirical Bayes estimates β_{0j}^{EB} must be biased towards the overall estimate γ_{00}. They are *shrunken* towards the average value γ_{00}. For that reason, the empirical Bayes estimators are also referred to as shrinkage estimators. Figure 2.7 presents boxplots for the OLS and EB estimates of the intercept and the extraversion regression slopes in the model without the cross-level interaction (model M1A in Table 2.3). It is clear that the OLS estimates have a higher variability.

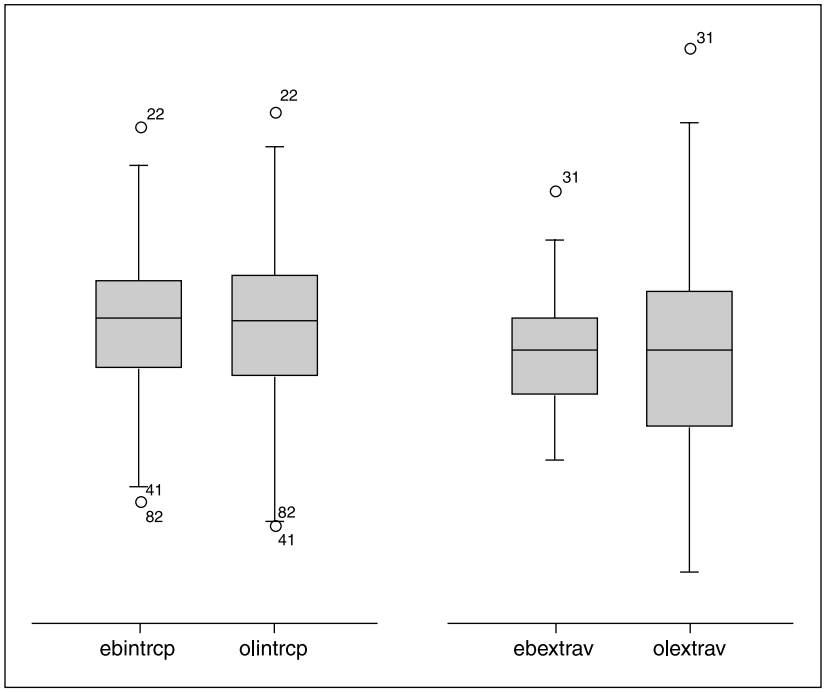

Figure 2.7 OLS and EB estimates for intercept and slope.

Although the empirical Bayes or shrinkage estimators are biased, they are also in general closer to the (unknown) values of β_{0j} (Bryk & Raudenbush, 1992, p. 40). If the regression model includes a group-level model, the shrinkage estimators are conditional on the group-level model. The advantages of shrinkage estimators remain, provided that the group-level model is well specified (Bryk & Raudenbush, 1992, p. 80). This is especially important if the estimated coefficients are used to describe specific groups. For instance, we can use estimates for the intercepts of the schools to rank

them on their average outcome. If this is used as an indicator of the quality of schools, the shrinkage estimators introduce a bias, because high scoring schools will be presented too negatively, and low scoring schools will be presented too positively. This is offset by the advantage of having a smaller standard error (Carlin & Louis, 1996; Lindley & Smith, 1972). Bryk and Raudenbush discuss this problem in an example involving the effectiveness of organizations (Bryk & Raudenbush, 1992, Chapter 5); see also the cautionary points made by Raudenbush and Willms (1991) and Snijders and Bosker (1999, pp. 58–63). All emphasize that the higher precision of the empirical Bayes residuals is achieved at the expense of a certain bias. The bias is largest when we inspect groups that are both small and far removed from the overall mean. In such cases, inspecting residuals should be supplemented with other procedures, such as comparing error bars for all schools (Goldstein & Healy, 1995). Error bars are illustrated in this chapter in Figure 2.4.

2.4 THREE- AND MORE-LEVEL REGRESSION MODELS

2.4.1 Multiple-level models

In principle, the extension of the two-level regression model to three and more levels is straightforward. There is an outcome variable at the first, the lowest level. In addition, there may be explanatory variables at all available levels. The problem is that three- and more-level models can become complicated very fast. In addition to the usual fixed regression coefficients, we must entertain the possibility that regression coefficients for first-level explanatory variables may vary across units of both the second and the third level. Regression coefficients for second-level explanatory variables may vary across units of the third level. To explain such variation, we must include cross-level interactions in the model. Regression slopes for the cross-level interaction between first-level and second-level variables may themselves vary across third-level units. To explain such variation, we need a three-way interaction involving variables at all three levels.

The equations for such models are complicated, especially when we do not use the more compact summation notation but write out the complete single-equation version of the model in an algebraic format (for a note on notation see section 2.5).

The resulting models are not only difficult to follow from a conceptual point of view; they may also be difficult to estimate in practice. The number of estimated parameters is considerable, and at the same time the highest-level sample size tends to become relatively smaller. As DiPrete and Forristal (1994, p. 349) put it, the imagination of the researchers '. . . can easily outrun the capacity of the data, the computer, and current optimization techniques to provide robust estimates.'

Nevertheless, three- and more-level models have their place in multilevel

analysis. Intuitively, three-level structures such as pupils in classes in schools, or respondents nested within households, nested within regions, appear to be both conceptually and empirically manageable. If the lowest level is repeated measures over time, having repeated measures on pupils nested within schools again does not appear to be overly complicated. In such cases, the solution for the conceptual and statistical problems mentioned is to keep models reasonably small. Especially specification of the higher-level variances and covariances should be driven by theoretical considerations. A higher-level variance for a specific regression coefficient implies that this regression coefficient is assumed to vary across units at that level. A higher-level covariance between two specific regression coefficients implies that these regression coefficients are assumed to covary across units at that level. Especially when models become large and complicated, it is advisable to avoid higher-order interactions, and to include in the random part only those elements for which there is strong theoretical or empirical justification. This implies that an exhaustive search for second-order and higher-order interactions is not a good idea. In general, we should look for higher-order interactions only if there is strong theoretical justification for their importance, or if an unusually large variance component for a regression slope calls for explanation. For the random part of the model, there are usually more convincing theoretical reasons for the higher-level variance components than for the covariance components. Especially if the covariances are small and insignificant, analysts sometimes do not include all possible covariances in the model. This is defensible, with some exceptions. First, it is recommended that the covariances between the intercept and the random slopes are always included. Second, it is recommended to include covariances corresponding to slopes of dummy variables belonging to the same categorical variable, and for variables that are involved in an interaction or belong to the same polynomial expression (Longford, 1990, pp. 79–80).

2.4.2 Intraclass correlations in three-level models

In a two-level model, the intraclass correlation is calculated in the intercept-only model using equation 2.9, which is repeated below:

$$\rho = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_e}. \qquad \text{(2.9, repeated)}$$

The intraclass correlation is an indication of the proportion of variance at the second level, and it can also be interpreted as the expected (population) correlation between two randomly chosen individuals within the same group.

If we have a three-level model, for instance pupils nested within classes, nested within schools, there are several ways to calculate the intraclass correlation. First, we

estimate an intercept-only model for the three-level data, for which the single-equation model can be written as follows:

$$Y_{ijk} = \gamma_{000} + v_{0k} + u_{0jk} + e_{ijk}. \tag{2.15}$$

The variances at the first, second, and third level are respectively σ_e^2, $\sigma_{u_0}^2$, and $\sigma_{v_0}^2$. The first method (see Davis & Scott, 1995) defines the intraclass correlations at the class and school level as:

$$\rho_{class} = \frac{\sigma_{u_0}^2}{\sigma_{v_0}^2 + \sigma_{u_0}^2 + \sigma_e^2}, \tag{2.16}$$

and:

$$\rho_{school} = \frac{\sigma_{v_0}^2}{\sigma_{v_0}^2 + \sigma_{u_0}^2 + \sigma_e^2}. \tag{2.17}$$

The second method (see Siddiqui, Hedeker, Flay, & Hu, 1996) defines the intraclass correlations at the class and school level as:

$$\rho_{class} = \frac{\sigma_{v_0}^2 + \sigma_{u_0}^2}{\sigma_{v_0}^2 + \sigma_{u_0}^2 + \sigma_e^2}, \tag{2.18}$$

and:

$$\rho_{school} = \frac{\sigma_{v_0}^2}{\sigma_{v_0}^2 + \sigma_{u_0}^2 + \sigma_e^2}. \tag{2.19}$$

Actually, both methods are correct (Algina, 2000). The first method identifies the proportion of variance at the class and school level. This should be used if we are interested in a decomposition of the variance across the available levels, or if we are interested in how much variance is explained at each level (a topic discussed in section 4.5). The second method represents an estimate of the expected (population) correlation between two randomly chosen elements in the same group. So ρ_{class} as calculated in equation 2.18 is the expected correlation between two pupils within the same class, and it correctly takes into account that two pupils who are in the same class must by definition also be in the same school. For this reason, the variance components for classes and schools must both be in the numerator of equation 2.18. If the two sets of estimates are different, which may happen if the amount of variance at the school level

is large, there is no contradiction involved. Both sets of equations express two different aspects of the data, which happen to coincide when there are only two levels.

2.4.3 An example of a three-level model

The data in this example are from a hypothetical study on stress in hospitals. The data are from nurses working in wards nested within hospitals. In each of 25 hospitals, four wards are selected and randomly assigned to an experimental and control condition. In the experimental condition, a training program is offered to all nurses to cope with job-related stress. After the program is completed, a sample of about 10 nurses from each ward is given a test that measures job-related stress. Additional variables are: nurse age (years), nurse experience (years), nurse gender (0 = male, 1 = female), type of ward (0 = general care, 1 = special care), and hospital size (0 = small, 1 = medium, 2 = large).

This is an example of an experiment where the experimental intervention is carried out at the group level. In biomedical research this design is known as a cluster randomized trial. They are quite common also in educational and organizational research, where entire classes or schools are assigned to experimental and control conditions. Since the design variable Experimental versus Control group (*ExpCon*) is manipulated at the second (ward) level, we can study whether the experimental effect is different in different hospitals, by defining the regression coefficient for the *ExpCon* variable as random at the hospital level.

In this example, the variable *ExpCon* is of main interest, and the other variables are covariates. Their function is to control for differences between the groups, which should be small given that randomization is used, and to explain variance in the outcome variable stress. To the extent that they are successful in explaining variance, the power of the test for the effect of *ExpCon* will be increased. Therefore, although logically we can test if explanatory variables at the first level have random coefficients at the second or third level, and if explanatory variables at the second level have random coefficients at the third level, these possibilities are not pursued. We do test a model with a random coefficient for *ExpCon* at the third level, where there turns out to be significant slope variation. This varying slope can be predicted by adding a cross-level interaction between the variables *ExpCon* and *HospSize*. In view of this interaction, the variables *ExpCon* and *HospSize* have been centered on their overall mean. Table 2.5 presents the results for a series of models.

The equation for the first model, the intercept-only model, is:

$$stress_{ijk} = \gamma_{000} + v_{0k} + u_{0jk} + e_{ijk}. \tag{2.20}$$

This produces the variance estimates in the M0 column of Table 2.5. The proportion of variance (ICC) is .52 at the ward level, and .17 at the hospital level, calculated

Table 2.5 Models for stress in hospitals and wards

Model	M0: intercept only	M1: with predictors	M2: with random slope *ExpCon*	M3: with cross-level interaction
Fixed part	Coef. (s.e.)	Coef. (s.e.)	Coef. (s.e.)	Coef. (s.e.)
Intercept	5.00 (0.11)	5.50 (.12)	5.46 (.12)	5.50 (.11)
ExpCon		–0.70 (.12)	–0.70 (.18)	–0.50 (.11)
Age		0.02 (.002)	0.02 (.002)	0.02 (.002)
Gender		–0.45 (.04)	–0.46 (.04)	–0.46 (.04)
Experience		–0.06 (.004)	–0.06 (.004)	–0.06 (.004)
Ward type		0.05 (.12)	0.05 (.07)	0.05 (.07)
HospSize		0.46 (.12)	0.29 (.12)	–0.46 (.12)
ExpCon × HospSize				1.00 (.16)
Random part				
σ^2_{eijk}	0.30 (.01)	0.22 (.01)	0.22 (.01)	0.22 (.01)
σ^2_{u0jk}	0.49 (.09)	0.33 (.06)	0.11 (.03)	0.11 (.03)
σ^2_{v0k}	0.16 (.09)	0.10 (.05)	0.166 (.06)	0.15 (.05)
σ^2_{u1j}			0.66 (.22)	0.18 (.09)
Deviance	1942.4	1604.4	1574.2	1550.8

following equations 2.16 and 2.17. The nurse-level and the ward-level variances are evidently significant. The test statistic for the hospital-level variance is $Z = 0.162/0.0852 = 1.901$, which produces a one-sided p-value of .029. The hospital-level variance is significant at the 5% level. The sequence of models in Table 2.5 shows that all predictor variables have a significant effect, except the ward type, and that the experimental intervention significantly lowers stress. The experimental effect varies across hospitals, and a large part of this variation can be explained by hospital size; in large hospitals the experimental effect is smaller.

2.5 A NOTE ABOUT NOTATION AND SOFTWARE

2.5.1 Notation

In general, there will be more than one explanatory variable at the lowest level and more than one explanatory variable at the highest level. Assume that we have P

explanatory variables X at the lowest level, indicated by the subscript p ($p = 1 \ldots P$). Likewise, we have Q explanatory variables Z at the highest level, indicated by the subscript q ($q = 1 \ldots Q$). Then, equation 2.5 becomes the more general equation:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{0q}Z_{qj} + \gamma_{pq}Z_{qj}X_{pij} + u_{pj}X_{pij} + u_{0j} + e_{ij}. \tag{2.21}$$

Using summation notation, we can express the same equation as:

$$Y_{ij} = \gamma_{00} + \sum_{p} \gamma_{p0}X_{pij} + \sum_{q} \gamma_{0q}Z_{qj} + \sum_{p}\sum_{q} \gamma_{pq}X_{pij}Z_{qj} + \sum_{p} u_{pj}X_{pij} + u_{0j} + e_{ij}. \tag{2.22}$$

The errors at the lowest level e_{ij} are assumed to have a normal distribution with a mean of zero and a common variance σ_e^2 in all groups. The u-terms u_{0j} and u_{pj} are the residual error terms at the highest level. They are assumed to be independent from the errors e_{ij} at the individual level, and to have a multivariate normal distribution with means of zero. The variance of the residual errors u_{0j} is the variance of the intercepts between the groups, symbolized by $\sigma_{u_0}^2$. The variances of the residual errors u_{pj} are the variances of the slopes between the groups, symbolized by $\sigma_{u_p}^2$. The *covariances* between the residual error terms $\sigma_{u_{pp'}}$ are generally not assumed to be zero; they are collected in the higher-level variance/covariance matrix Ω.[8]

Note that in equation 2.22, γ_{00}, the regression coefficient for the intercept, is not associated with an explanatory variable. We can expand the equation by providing an explanatory variable that is a constant equal to one for all observed units. This yields the equation:

$$Y_{ij} = \gamma_{p0}X_{pij} + \gamma_{pq}Z_{qj}X_{pij} + u_{pj}X_{pij} + e_{ij} \tag{2.23}$$

where $X_{0ij} = 1$, and $p = 0 \ldots P$. Equation 2.23 makes clear that the intercept is a regression coefficient, just like the other regression coefficients in the equation. Some multilevel software, for instance HLM (Raudenbush, Bryk, Cheong, & Congdon, 2004), puts the intercept variable $X_0 = 1$ in the regression equation by default. Other multilevel software, for instance MLwiN (Rasbash, Steele, Browne, & Goldstein, 2009), requires that the analyst includes a variable in the data set that equals one in all cases, which must be added explicitly to the regression equation. In some cases, being able to eliminate the intercept term from the regression equation is a convenient feature.

[8] We may attach a subscript to Ω to indicate to which level it belongs. As long as there is no risk of confusion, the simpler notation without the subscript is used.

Equation 2.23 can be made very general if we let **X** be the matrix of all explanatory variables in the fixed part, symbolize the residual errors at all levels by $\mathbf{u}^{(l)}$ with l denoting the level, and associate all error components with predictor variables **Z**, which may or may not be equal to the **X**. This produces the very general matrix formula $\mathbf{Y} = \mathbf{X\beta} + \mathbf{Z}^{(l)}\mathbf{u}^{(l)}$ (see Goldstein, 1995, Appendix 2.1). Since this book is more about applications than about mathematical statistics, it generally uses the algebraic notation, except when multivariate procedures such as structural equation modeling are discussed.

The notation used in this book is close to the notation used by Goldstein (1987, 2003), Hox (1995), and Kreft and de Leeuw (1998). The most important difference is that these authors indicate the higher-level variance by σ_{00} instead of our $\sigma^2_{u_0}$. The logic is that, if σ_{01} indicates the covariance between variables 0 and 1, then σ_{00} is the covariance of variable 0 with itself, which is its variance. Bryk and Raudenbush (1992) and Snijders and Bosker (1999) use a different notation; they denote the lowest-level error terms by r_{ij}, and the higher-level error terms by u_j. The lowest-level variance is σ^2 in their notation. The higher-level variances and covariances are indicated by the Greek letter *tau*; for instance, the intercept variance is given by τ_{00}. The τ_{pp} are collected in the matrix TAU, symbolized as T. The HLM program and manual in part use a different notation, for instance when discussing longitudinal and three-level models.

In models with more than two levels, two different notational systems are used. One approach is to use different Greek characters for the regression coefficients at different levels, and different (Greek or Latin) characters for the variance terms at different levels. With many levels, this becomes cumbersome, and it is simpler to use the same character, say β for the regression slopes and u for the residual variance terms, and let the number of subscripts indicate to which level these belong.

2.5.2 Software

Multilevel models can be formulated in two ways: (1) by presenting separate equations for each of the levels, and (2) by combining all equations by substitution into a single model equation. The software HLM (Raudenbush et al., 2004) requires specification of the separate equations at each available level, but it can also show the single-equation version. Most other software, for example MLwiN (Rasbash et al., 2009), SAS Proc Mixed (Littell et al., 1996), SPSS Command Mixed (Norusis, 2005), uses the single-equation representation. Both representations have their advantages and disadvantages. The separate-equation representation has the advantage that it is always clear how the model is built up. The disadvantage is that it hides from view that modeling regression slopes by other variables results in adding an interaction to the model. As will be explained in Chapter 4, estimating and interpreting interactions correctly requires careful thinking. On the other hand, while the single-equation repre-

sentation makes the existence of interactions obvious, it conceals the role of the complicated error components that are created by modeling varying slopes. In practice, to keep track of the model, it is recommended to start by writing the separate equations for the separate levels, and to use substitution to arrive at the single-equation representation.

To take a quote from Singer's excellent introduction to using SAS Proc Mixed for multilevel modeling (Singer, 1998, p. 350): 'Statistical software does not a statistician make. That said, without software, few statisticians and even fewer empirical researchers would fit the kinds of sophisticated models being promulgated today.' Indeed, software does not make a statistician, but the advent of powerful and user-friendly software for multilevel modeling has had a large impact in research fields as diverse as education, organizational research, demography, epidemiology, and medicine. This book focuses on the conceptual and statistical issues that arise in multilevel modeling of complex data structures. It assumes that researchers who apply these techniques have access to and familiarity with *some* software that can estimate these models. Software is mentioned in various places, especially when a technique is discussed that requires specific software features or is only available in a specific program.

Since statistical software evolves rapidly, with new versions of the software coming out much faster than new editions of general handbooks such as this, I do not discuss software setups or output in detail. As a result, this book is more about the possibilities offered by the various techniques than about how these things can be done in a specific software package. The techniques are explained using analyses on small but realistic data sets, with examples of how the results could be presented and discussed. At the same time, if the analysis requires that the software used have some specific capacities, these are pointed out. This should enable interested readers to determine whether their software meets these requirements, and assist them in working out the software setups for their favorite package.

In addition to the relevant program manuals, several software programs have been discussed in introductory articles. Using SAS Proc Mixed for multilevel and longitudinal data is discussed by Singer (1998). Peugh and Enders (2005) discuss SPSS Mixed using Singer's examples. Both Arnold (1992), and Heck and Thomas (2009) discuss multilevel modeling using HLM and Mplus as the software tool. Sullivan, Dukes, and Losina (1999) discuss HLM and SAS Proc Mixed. West, Welch, and Gatecki (2007) present a series of multilevel analyses using SAS, SPSS, R, Stata, and HLM. Finally, the multilevel modeling program at the University of Bristol maintains a multilevel homepage that contains a series of software reviews. The homepage for this book (on www.joophox.net) contains links to these and other multilevel resources.

The data used in the various examples are described in Appendix A, and are all available through the Internet.

3

Estimation and Hypothesis Testing in Multilevel Regression

The usual method to estimate the values of the regression coefficients and the intercept and slope variances is the maximum likelihood method. This chapter gives a non-technical explanation of this estimation method, to enable analysts to make informed decisions on the estimation options presented by present software. Some alternatives to maximum likelihood estimation are briefly discussed. Recent developments, such as bootstrapping and Bayesian estimation methods, are also briefly introduced in this chapter. In addition, these are explained in more detail in Chapter 13. Finally, this chapter describes some procedures that can be used to test hypotheses about specific parameters.

3.1 WHICH ESTIMATION METHOD?

Estimation of parameters (regression coefficients and variance components) in multilevel modeling is mostly done by the maximum likelihood method. The maximum likelihood (ML) method is a general estimation procedure, which produces estimates for the population parameters that maximize the probability (produce the 'maximum likelihood') of observing the data that are actually observed, given the model (see Eliason, 1993). Other estimation methods that have been used in multilevel modeling are generalized least squares (GLS), generalized estimating equations (GEE), and Bayesian methods such as Markov chain Monte Carlo (MCMC). Bootstrapping methods (see Mooney & Duval, 1993) can be used to improve the parameter estimates and the standard errors. In this section, I will discuss these methods briefly.

3.1.1 Maximum likelihood

Maximum likelihood (ML) is the most commonly used estimation method in multilevel modeling. An advantage of the maximum likelihood estimation method is that it is generally robust, and produces estimates that are asymptotically efficient and consistent. With large samples, ML estimates are usually robust against mild violations of the assumptions, such as having non-normal errors. Maximum likelihood estimation proceeds by maximizing a function called the likelihood function. Two different

likelihood functions are used in multilevel regression modeling. One is full maximum likelihood (FML); in this method, both the regression coefficients and the variance components are included in the likelihood function. The other estimation method is restricted maximum likelihood (RML); here only the variance components are included in the likelihood function, and the regression coefficients are estimated in a second estimation step. Both methods produce parameter estimates with associated standard errors and an overall model *deviance*, which is a function of the likelihood. FML treats the regression coefficients as fixed but unknown quantities when the variance components are estimated, but does not take into account the degrees of freedom lost by estimating the fixed effects. RML estimates the variance components after removing the fixed effects from the model (see Searle, Casella, & McCulloch, 1992, Chapter 6). As a result, FML estimates of the variance components are biased; they are generally too small. RML estimates have less bias (Longford, 1993). RML also has the property that if the groups are balanced (have equal group sizes), the RML estimates are equivalent to analysis of variance (ANOVA) estimates, which are optimal (Searle et al., 1992, p. 254). Since RML is more realistic, it should, in theory, lead to better estimates, especially when the number of groups is small (Bryk & Raudenbush, 1992; Longford, 1993). In practice, the differences between the two methods are usually small (see Hox, 1998; Kreft & de Leeuw, 1998). For example, if we compare the FML estimates for the intercept-only model for the popularity data in Table 2.1 with the corresponding RML estimates, the only difference within two decimals is the intercept variance at level 2. FML estimates this as 0.69, and RML as 0.70. The size of this difference is absolutely trivial. If nontrivial differences are found, the RML method usually performs better (Browne, 1998). FML still continues to be used, because it has two advantages over RML. First, the computations are generally easier, and second, since the regression coefficients are included in the likelihood function, an overall chi-square test based on the likelihood can be used to compare two models that differ in the fixed part (the regression coefficients). With RML, only differences in the random part (the variance components) can be compared with this test. Most tables in this book have been produced using FML estimation, if RML is used this is explicitly stated in the text.

Computing the maximum likelihood estimates requires an *iterative* procedure. At the start, the computer program generates reasonable starting values for the various parameters (in multilevel regression analysis these are usually based on single-level regression estimates). In the next step, an ingenious computation procedure tries to improve on the starting values, to produce better estimates. This second step is repeated (iterated) many times. After each iteration, the program inspects how much the estimates have actually changed compared to the previous step. If the changes are very small, the program concludes that the estimation procedure has *converged* and that it is finished. Using multilevel software, we generally take the computational details for

granted. However, computational problems do sometimes occur. A problem common to programs using an iterative maximum likelihood procedure is that the iterative process is not always *guaranteed* to stop. There are models and data sets for which the program may go through an endless sequence of iterations, which can only be ended by stopping the program. Because of this, most programs set a built-in limit for the maximum number of iterations. If convergence is not reached within this limit, the computations can be repeated with a higher limit. If the computations do not converge after an extremely large number of iterations, we suspect that they may never converge.[1] The problem is how one should interpret a model that does not converge. The usual interpretation is that a model for which convergence cannot be reached is a bad model, using the simple argument that if estimates cannot be found, this disqualifies the model. However, the problem may also lie with the data. Especially with small samples, the estimation procedure may fail even if the model is valid. In addition, it is even possible that, if only we had a better computer algorithm, or better starting values, we could find acceptable estimates. Still, experience shows that if a program does not converge with a data set of reasonable size, the problem often is a badly misspecified model. In multilevel analysis, non-convergence often occurs when we try to estimate too many random (variance) components that are actually close or equal to zero. The solution is to simplify the model by leaving out some random components; often the estimated values from the non-converged solution provide an indication of which random components can be omitted.

3.1.2 Generalized least squares

Generalized least squares (GLS) is an extension of the standard estimation ordinary least squares (OLS) method that allows for heterogeneity and observations that differ in sampling variance. GLS estimates approximate ML estimates, and they are asymptotically equivalent. Asymptotic equivalence means that in very large samples they are in practice indistinguishable. Goldstein (2003, p. 21) notes that 'expected GLS' estimates can be obtained from a maximum likelihood procedure by restricting the number of iterations to one. Since GLS estimates are obviously faster to compute than full ML estimates, they can be used as a stand-in for ML estimates in computationally intensive procedures such as extremely large data sets, or when bootstrapping is used. They can also be used when ML procedures fail to converge; inspecting the GLS results may help to diagnose the problem. Furthermore, since GLS estimates are respectable statistical estimates in their own right, in such situations one can report the GLS estimates instead of the more usual ML estimates. However, simulation research

[1] Some programs allow the analyst to monitor the iterations, to observe whether the computations are going somewhere, or are just moving back and forth without improving the likelihood function.

shows that, in general, GLS estimates are less efficient, and the GLS-derived standard errors are rather inaccurate (see Hox, 1998; Kreft, 1996; van der Leeden, Meijer, & Busing, 2008). Therefore, in general, ML estimation should be preferred.

3.1.3 Generalized estimating equations

The generalized estimating equations method (GEE, see Liang & Zeger, 1986) estimates the variances and covariances in the random part of the multilevel model directly from the residuals, which makes them faster to compute than full ML estimates. Typically, the dependences in the multilevel data are accounted for by a very simple model, represented by a *working correlation matrix*. For individuals within groups, the simplest assumption is that the respondents within the same group all have the same correlation. For repeated measures, a simple autocorrelation structure is usually assumed. After the estimates for the variance components are obtained, GLS is used to estimate the fixed regression coefficients. Robust standard errors are generally used to counteract the approximate estimation of the random structure. For non-normal data this results in a *population-average model*, where the emphasis is on estimating average population effects and not on modeling individual and group differences. Raudenbush and Bryk (2002) describe the multilevel *unit-specific* model (usually based on ML estimation) as a model that aims to model the effect of predictor variables while controlling statistically for other predictor variables at different levels, plus the random effects in the model. In contrast, the *population-average* model (usually based on GEE estimation) controls for the other predictor variables, but not for the random effects. When a nonlinear model is estimated, the GEE estimates are different from the ML estimates. For example, in an intercept-only logistic regression model the average probability in the population of repeating a class can be calculated from the population-average estimate of the intercept. The unit-specific intercept can in general not be used to calculate this probability. If the interest is group-level variation, for instance in modeling the differences in the level 1 effects using level 2 variables, the unit-specific model is appropriate. If we are only interested in population estimates of the average effect of level 1 variables, for instance in the difference between boys and girls nationwide in the probability of repeating a class, the population-average model is appropriate. For a further discussion, I refer to Zeger, Liang, and Albert (1988) and Hu, Goldberg, Hedeker, Flay, and Pentz (1998).

According to Goldstein (2003) and Raudenbush and Bryk (2002), GEE estimates are less efficient than full ML estimates, but they make weaker assumptions about the structure of the random part of the multilevel model. If the model for the random part is correctly specified, ML estimators are more efficient, and the model-based (ML) standard errors are generally smaller than the GEE-based robust standard errors. If the model for the random part is incorrect, the GEE-based estimates and

robust standard errors are still consistent. So, provided the sample size is reasonably large, GEE estimators are robust against misspecification of the random part of the model, including violations of the normality assumption. A drawback of the GEE approach is that it only approximates the random effects structure, and therefore the random effects cannot be analyzed in detail. So, most software will estimate a full unstructured covariance matrix for the random part, which makes it impossible to estimate random effects for the intercept or slopes. Given the general robustness of ML methods, it is preferable to use ML methods when these are available, and to use robust estimators or bootstrap corrections when there is serious doubt about the assumptions of the ML method. Robust estimators, which are related to GEE estimators (Burton, Gurrin, & Sly, 1998), and bootstrapping are treated in more detail in Chapter 13 of this book.

3.1.4 Bootstrapping

In bootstrapping, random samples are repeatedly drawn with replacement from the observed data. In each of these random samples, the model parameters are estimated, generally using either FML or RML maximum likelihood estimation. This process is repeated b times. For each model parameter, this results in a set of b parameter estimates. The variance of these b estimates is used as an indicator of the sampling variance associated with the parameter estimate obtained from the full sample. Since the bootstrap samples are obtained by resampling from the total sample, bootstrapping falls under the general term of resampling methods (see Good, 1999). Bootstrapping can be used to improve both the point estimates and the standard errors. Typically, at least 1000 bootstrap samples are needed for sufficient accuracy. This makes the method computationally demanding, but less so than the Bayesian methods treated in the next section. Since bootstrapping has its own complications, it is discussed in more detail in Chapter 13. If we execute a bootstrap estimation for our example data, the results are almost identical to the asymptotic FML results reported in Table 2.2. The estimates differ by 0.01 at most, which is a completely trivial difference. Bootstrap estimates are most attractive when we have reasons to suspect the asymptotic results, for example because we have a small sample size, or because we have non-normal data.

3.1.5 Bayesian methods

In Bayesian statistics, we express our uncertainty about the population values of the model parameters by assigning to them a distribution of possible values. This distribution is called the *prior* distribution, because it is specified independently from the data. The prior distribution is combined with the likelihood of the data to produce a *posterior* distribution, which describes our uncertainty about the population values

after observing the data. Typically, the variance of the posterior distribution is smaller than the variance of the prior distribution, which means that observing the data has reduced our uncertainty about the possible population values. For the prior distribution, we have a fundamental choice between using an informative prior and using an uninformative prior. An informative prior is a peaked distribution with a small variance, which expresses a strong belief about the unknown population parameter. An informative prior will, of course, strongly influence the posterior distribution, and hence our conclusions. For this reason, many statisticians prefer an uninformative or diffuse prior, which has very little influence on the posterior, and only serves to produce the posterior. An example of an uninformative prior is the uniform distribution, which simply states that the unknown parameter value is a number between minus and plus infinity, with all values equally likely.

If the posterior distribution has a mathematically simple form, for instance a normal distribution, we can use this distribution to produce a point estimate and a confidence interval for the population parameter. However, in complex multivariate models, the posterior is generally a complicated multivariate distribution, which makes it difficult to use it directly to produce parameter estimates and confidence intervals. Therefore, simulation techniques are used to generate random samples from the posterior distribution. The simulated posterior distribution is then used to provide a point estimate (typically the mode or median of the simulated values) and a confidence interval.

Bayesian methods can provide accurate estimates of the parameters and the uncertainty associated with them (Goldstein, 2003). However, they are computationally demanding, and the simulation procedure must be monitored to insure that it is working properly. Bayesian estimation methods are treated in more detail in Chapter 13.

3.2 SIGNIFICANCE TESTING AND CONFIDENCE INTERVALS

This section discusses procedures for testing significance and constructing confidence intervals for the regression coefficients and variance components. There is also a global test based on the likelihood that can be used to compare two nested models. Finally, contrasts of regression coefficients or variances are described, and constraints that can be used to constrain regression coefficients or variance components to specific values.

3.2.1 Testing regression coefficients and variance components

Maximum likelihood estimation produces parameter estimates and corresponding standard errors. These can be used to carry out a significance test of the form $Z =$

(*estimate*)/(*standard error of estimate*), where Z is referred to the standard normal distribution. This test is known as the *Wald test* (Wald, 1943). The standard errors are asymptotic, which means that they are valid for large samples. As usual, it is not precisely known when a sample is large enough to give confidence about the precision of the estimates. Simulation research suggests that for accurate standard errors for level 2 variances, a relatively large level 2 sample size is needed. For instance, simulations by van der Leeden, Busing, and Meijer (1997) suggest that with fewer than 100 groups, ML estimates of variances and their standard errors are not very accurate. In ordinary regression analysis, a rule of thumb is to require $104 + p$ observations if the interest is in estimating and interpreting regression coefficients, where p is the number of explanatory variables (Green, 1991). If the interest is in interpreting (explained) variance, the rule of thumb is to require $50 + 8p$ observations. In multilevel regression, the relevant sample size for higher-level coefficients and variance components is the number of groups, which is often not very large. Green's rule of thumb and van der Leeden et al.'s simulation results agree on a preferred group-level sample size of at least 100. Additional simulation research (Maas & Hox, 2005) suggests that if the interest lies mostly in the fixed part of the model, far fewer groups are sufficient, especially for the lowest-level regression coefficients. The issue of the sample sizes needed to produce accurate estimates and standard errors is taken up in more detail in Chapter 12.

It should be noted that the p-values and confidence intervals produced by the program HLM (Raudenbush, Yang, & Yosef, 2000) differ from those obtained from most other programs. As part of their output, most multilevel analysis programs produce parameter estimates and asymptotic standard errors for these estimates, all obtained from the maximum likelihood estimation procedure. The usual significance test is the Wald test, with Z evaluated against the standard normal distribution. Bryk and Raudenbush (1992, p. 50), referring to a simulation study by Fotiu (1989), argue that for the fixed effects it is better to refer this ratio to a t-distribution on $J - p - 1$ degrees of freedom, where J is the number of second level units, and p is the total number of explanatory variables in the model. The p-values produced by the program HLM (Raudenbush, Bryk, Cheong, & Congdon, 2004) are based on these tests rather than on the more common Wald tests. When the number of groups J is large, the difference between the asymptotic Wald test and the alternative Student t-test is very small. However, when the number of groups is small, the differences may become important. Since referring the result of the Z-test on the regression coefficients to a Student t-distribution is conservative, this procedure should provide a better protection against type I errors. A better choice for the degrees of freedom in multilevel models is provided by the Satterthwaite approximation (Sattertwaithe, 1946). This estimates the number of degrees of freedom using the values of the residual variances. Simulation research (Manor & Zucker, 2004) shows that the Satterthwaite approximation works better than the Wald test when the sample size is small (e.g., smaller than 30).

Several authors (e.g., Berkhof & Snijders, 2001; Raudenbush & Bryk, 2002) argue that the Z-test is not appropriate for the variances, because it assumes a normal distribution, whereas the sampling distribution of variances is skewed, especially when the variance is small. Especially if we have both a small sample of groups and a variance component close to zero, the distribution of the Wald statistic is clearly non-normal. Raudenbush and Bryk propose to test variance components using a chi-square test on the residuals. This chi-square is computed by:

$$\chi^2 = \sum (\hat{\beta}_j - \beta)^2 / \hat{V}_j, \tag{3.1}$$

where $\hat{\beta}_j$ is the OLS estimate of a regression coefficient computed separately in group j, β its overall estimate, and $\hat{V}_j$ its estimated sampling variance in group j. The number of degrees of freedom is given by $df = J - p - 1$, where J is the number of second-level units, and p is the total number of explanatory variables in the model. Groups that have a small number of cases are passed over in this test, because their OLS estimates are not sufficiently accurate.

Simulation studies on the Wald test for a variance component (van der Leeden et al., 1997) and the alternative chi-square test (Harwell, 1997; Sánchez-Meca & Marín-Martínez, 1997) suggest that with small numbers of groups, both tests suffer from a very low power. The test that compares a model with and without the parameters under consideration, using the chi-square model test described in the next section, is generally better (Berkhof & Snijders, 2001; Goldstein, 2003). Only if the likelihood is determined with a low precision, which is the case for some approaches to modeling non-normal data, is the Wald test preferred. Note that if the Wald test is used to test a variance component, a one-sided test is the appropriate one.

3.2.2 Comparing nested models

From the likelihood function we can calculate a statistic called the *deviance* that indicates how well the model fits the data. The deviance is defined as $-2 \times \ln$ (*Likelihood*), where *Likelihood* is the value of the likelihood function at convergence, and is the natural logarithm. In general, models with a lower deviance fit better than models with a higher deviance. If two models are *nested*, which means that a specific model can be derived from a more general model by removing parameters from the general model, we can compare them statistically using their deviances. The difference of the deviances for two nested models has a chi-square distribution, with degrees of freedom equal to the difference in the number of parameters estimated in the two models. This can be used to perform a formal chi-square test to test whether the more general model fits significantly better than the simpler model. The deviance difference test is also

referred to as the likelihood ratio test, since the ratio of two likelihoods is compared by looking at the difference of their logarithms.

The chi-square test of the deviances can be used to good effect to explore the importance of random effects, by comparing a model that contains these effects with a model that excludes them.

Table 3.1 presents two models for the pupil popularity data used as an example in Chapter 2. The first model contains only an intercept. The second model adds two pupil-level variables and a teacher-level variable, with the pupil-level variable extraversion having random slopes at the second (class) level. To test the second-level variance component σ^2_{u0} using the deviance difference test, we remove it from model M0. The resulting model (not presented in Table 3.1) produces a deviance of 6970.4, and the deviance difference is 642.9. Since the modified model estimates one parameter less, this is referred to the chi-square distribution with one degree of freedom. The result is obviously significant.

Table 3.1 Intercept-only model and model with explanatory variables

Model	M0: intercept only	M1: with predictors
Fixed part	Coefficient (s.e.)	Coefficient (s.e.)
Intercept	5.08 (.09)	0.74 (.20)
Pupil gender		1.25 (.04)
Pupil extraversion		0.45 (.02)
Teacher experience		0.09 (.01)
Random part		
σ^2_e	1.22 (.04)	0.55 (.02)
σ^2_{u0}	0.69 (.11)	1.28 (.28)
σ_{u02}		–0.18 (.05)
σ^2_{u2}		0.03 (.008)
Deviance	6327.5	4812.8

The variance of the regression coefficient for pupil gender is estimated as zero, and therefore it is removed from the model. A formal test is not necessary. In model M1 in Table 3.1 this variable is treated as fixed: no variance component is estimated. To test the significance of the variance of the extraversion slopes, we must remove the variance parameter from the model. This presents us with a problem, since there is also a covariance parameter σ_{u02} associated with the extraversion slopes. If we remove both

the variance and the covariance parameter from the model, we are testing a combined hypothesis on two degrees of freedom. It is better to separate these tests. Some software (e.g., MLwiN) actually allows us to remove the variance of the slopes from the model but to retain the covariance parameter. This is a strange model, but for testing purposes it allows us to carry out a separate test on the variance parameter only. Other software (e.g., MLwiN, SPSS, SAS) allows the removal of the covariance parameter, while keeping the variance in the model. If we modify model M1 this way, the deviance increases to 4851.9. The difference is 39.1, which is a chi-square variate with one degree of freedom, and highly significant. If we modify the model further, by removing the slope variance, the deviance increases again to 4862.3. The difference with the previous model is 10.4, again with one degree of freedom, and it is highly significant.

Asymptotically, the Wald test and the test using the chi-square difference are equivalent. In practice, the Wald test and the chi-square difference test do not always lead to the same conclusion. If a variance component is tested, the chi-square difference test is clearly better, except when models are estimated where the likelihood function is only an approximation, as in the logistic models discussed in Chapter 6.

When the chi-square difference test is used to test a variance component, it should be noted that the standard application leads to a *p*-value that is too high. The reason is that the null-hypothesis of zero variance is on the boundary of the parameter space (all possible parameter values) since variances cannot be negative. If the null-hypothesis is true, there is a 50% chance of finding a positive variance, and a 50% chance of finding a negative variance. Negative variances are inadmissible, and the usual procedure is to change the negative estimate to zero. Thus, under the null-hypothesis the chi-square statistic has a mixture distribution of 50% zero and 50% chi-square with one degree of freedom. Therefore, the *p*-value from the chi-square difference test must be divided by two if a variance component is tested (Berkhof & Snijders, 2001). If we test a slope variance, and remove both the slope variance and the covariance from the model, the mixture is more complicated, because we have a mixture of 50% chi-square with one degree of freedom for the unconstrained intercept–slope covariance and 50% chi-square with two degrees of freedom for the covariance and the variance that is constrained to be non-negative (Verbeke & Molenberghs, 2000). The *p*-value for this mixture is calculated using $p = 0.5\,P\left(\chi_1^2 > C^2\right) + 0.5P\left(\chi_2^2 > C^2\right)$ where C^2 is the difference in the deviances of the model with and without the slope variance and intercept–slope covariance. Stoel, Galindo, Dolan, and van den Wittenboer (2006) discuss how to carry out such tests in general. If it is possible to remove the intercept–slope covariance from the model, it is possible to test the significance of the slope variance with a one degree of freedom test, and we can simply halve the *p*-value again. For the regression coefficients, the chi-square test (only in combination with FML estimation) is in general also superior. The reason is that the Wald test is to some degree sensitive to the parameterization of the model and the specific restrictions to be

tested (Davidson & MacKinnon, 1993, Chapter 13.5–13.6). The chi-square test is invariant under reparameterizations of the model. Since the Wald test is much more convenient, it is in practice used the most, especially for the fixed effects. Even so, if there is a discrepancy between the result of a chi-square difference test and the equivalent Wald test, the chi-square difference test is generally the preferred one.

LaHuis and Ferguson (2009) compare, among others, the chi-square deviance test and the chi-square residuals test described above. In their simulation, all tests controlled the type I error well, and the deviance difference test (dividing p by two, as described above) generally performed best in terms of power.

3.2.3 Comparing non-nested models

If the models to be compared are not nested models, the principle that models should be as simple as possible (theories and models should be parsimonious) indicates that we should generally keep the simpler model. A general fit index to compare the fit of statistical models is Akaike's Information Criterion, *AIC* (Akaike, 1987), which was developed to compare non-nested models, adjusting for the number of parameters estimated. The AIC for multilevel regression models is conveniently calculated from the deviance d, and the number of estimated parameters q:

$$AIC = d + 2q. \tag{3.2}$$

The AIC is a very general fit index that assumes that we are comparing models that are fit to the same data set, using the same estimation method. A fit index similar to the AIC is Schwarz's Bayesian Information Criterion, *BIC* (Schwarz, 1978), which is given by:

$$BIC = d + q \ln(N). \tag{3.3}$$

In multilevel modeling, the general equation 3.3 for the BIC is ambiguous, because it is unclear whether N refers to the first-level or the second-level sample size. What N means in equation 3.3 is differently chosen by different software. Most software uses the number of units at the highest level for the N. This makes sense when multilevel models are used for longitudinal data, where the highest level is often the subject level. Given the strong interest in multilevel modeling in contextual effects, choosing the highest-level sample size appears a sensible rule.

When the deviance goes down, indicating a better fit, both the AIC and the BIC also tend to go down. However, the AIC and the BIC both include a penalty function based on the number of estimated parameters q. As a result, when the number of estimated parameters goes up, the AIC and BIC tend to go up too. For most sample

sizes, the BIC places a larger penalty on complex models, which leads to a preference for smaller models. Since multilevel data have a different sample size at different levels, the AIC is more straightforward than the BIC, and is therefore the recommended choice. The AIC and BIC are typically used to compare a range of competing models, and the model(s) with the lowest AIC or BIC value is considered the most attractive. Both the AIC and the BIC have been shown to perform well, with a small advantage for the BIC (Haughton, Oud, & Jansen, 1997; Kieseppä, 2003). It should be noted that the AIC and BIC are based on the likelihood function. With FML estimation, the AIC and BIC can be used to compare models that differ either in the fixed part or in the random part. If RML estimation is used, it can only be used to compare models that differ in the random part. Since RML effectively partials out the fixed part, before the random part is estimated, the RML likelihood may still change if the fixed part is changed. Therefore, if likelihood-based procedures are used to compare models using RML estimation, the fixed part of the model should be kept constant. Most current software does not produce the AIC or BIC (SPSS and SAS do), but they can be calculated using the formulas given earlier. For an introductory discussion of the background of the AIC and the BIC see McCoach and Black (2008).

3.3 CONTRASTS AND CONSTRAINTS

It is possible to define a *contrast* for a set of regression coefficients or variance components. A contrast is a composite hypothesis on a set of parameters, for example, that they are equal, or that a set of parameters are all equal to zero. In general, a contrast has the form:

$$H_0\text{: } \mathbf{Cp} = \mathbf{k}, \tag{3.4}$$

where **C** is a contrast matrix that specifies the composite hypothesis, **p** is the vector of parameters that are involved, and **k** is a vector of contrast values. An example makes clear how this works. Table 3.1 presents the results of analyzing the popularity data. There were 2000 pupils in 100 classes. The intercept-only model lists two random parameters: the intercept variance at level 1, which is estimated as 1.22, and the intercept variance at level 2, which is estimated as 0.69. Suppose that we want to evaluate the null-hypothesis that these variance components are equal. Thus, the null-hypothesis is: $\sigma_e^2 - \sigma_u^2 = 0$. The parameter vector **p** is:

$$\mathbf{p} = \begin{Bmatrix} \sigma_e^2 \\ \sigma_{u0}^2 \end{Bmatrix}. \tag{3.5}$$

We can write the composite null-hypothesis as:

$$\{1, -1\} \begin{Bmatrix} \sigma_e^2 \\ \sigma_{u0}^2 \end{Bmatrix} = \{0\}, \tag{3.6}$$

or, in simple algebraic notation:

$$1 \times \sigma_e^2 - 1 \times \sigma_{u0}^2 = 0. \tag{3.7}$$

Contrasts are tested using an asymptotic Wald chi-square test (Goldstein, 2003; Raudenbush & Bryk, 2002). If we test the null-hypothesis represented by contrast 3.6 on the example data, we obtain a chi-square of 21.1 with one degree of freedom, and a *p*-value of < .001. The null-hypothesis of equal variances at both levels is rejected.

Sometimes we have a composite hypothesis that is more complicated, and we may require more than one contrast vector. For example, suppose that we want to test the null-hypothesis that all three of the regression coefficients in Table 3.1 are equal. The list of all fixed coefficients is: the intercept, the regression coefficients for pupil extraversion and gender, and the regression coefficient for teacher experience. Our composite null-hypothesis is H_0: $\gamma_1 = \gamma_2 = \gamma_3$. This composite null-hypothesis can be expressed by having two simple null-hypotheses: $\gamma_1 - \gamma_2 = 0$ and $\gamma_2 - \gamma_3 = 0$. In matrix formulation, replacing the gammas by the corresponding variable names, we have:

$$\begin{Bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{Bmatrix} \begin{Bmatrix} intercept \\ pup.\ extrav. \\ pup.\ gender \\ teach.\ exper. \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}. \tag{3.8}$$

Equation 3.8 represents the two contrasts given by $0 \times \gamma_{00} + 1 \times \gamma_{01} + (-1) \times \gamma_{02} + 0 \times \gamma_{10} = 0$ and $0 \times \gamma_{00} + 0 \times \gamma_{01} + 1 \times \gamma_{02} + (-1) \times \gamma_{10} = 0$, or more simply $\gamma_{01} = \gamma_{02}$ and $\gamma_{02} = \gamma_{10}$. If we test this composite null-hypothesis on the example data, we obtain a chi-square of 313.8 with two degrees of freedom (one for each row of the contrast matrix), and a *p*-value of < .001. The null hypothesis is clearly untenable.

In these two examples, the null-hypothesis is not very interesting. Contrast tests on parameters are theoretically informative when we can use them to compare regression coefficients or variance components that refer to similar variables. For instance, in a multivariate multilevel analysis, where we have two or more outcome variables, we can use contrasts to test if a predictor variable has the same effect on both outcome variables. Alternatively, if we have two comparable predictor variables, we can test if they both have the same effect on a single outcome variable. Contrast tests are also useful when categorical variables are represented by a set of dummies. Using complex

contrasts, we can replace a set of separate Wald tests by one single omnibus test on the full set of dummies.

Some software allows putting *constraints* on a set of parameters. Simple constraints are constraining a single parameter to be equal to zero or one. This is used, for instance, in multilevel models for meta-analysis, which is treated in Chapter 11. Complex constraints are similar to complex contrasts, and allow constraining a set of parameters to be equal. A useful application is in models with dummy variables that identify groups. We can use a constraint matrix to specify that the variances for the group identification must be equal. This is used in models for cross-classified data, which is treated in Chapter 9.

Constraints are similar to contrasts, but do not specify a value for the contrast. If we compare models with and without constraints using the deviance chi-square test, we have a test that is asymptotically identical to the equivalent composite test using contrasts. The composite contrast test on model parameters is a composite Wald chi-square test (see Goldstein, 2003). The global model comparison using deviances is a likelihood ratio test. With normal data, the global likelihood ratio test is generally considered more accurate, especially when variance components are tested. With non-normal data, the likelihood is often approximated, and then the composite Wald test is preferred.

4

Some Important Methodological and Statistical Issues

The multilevel regression model is more complicated than the standard single-level multiple regression model. One difference is the number of parameters, which is much larger in the multilevel model. This poses problems when models are fitted that have many parameters, and also in model exploration. Another difference is that multilevel models often contain interaction effects in the form of cross-level interactions. Interaction effects are tricky, and analysts should deal with them carefully. Finally, the multilevel model contains several different residual variances, and no single number can be interpreted as *the* amount of explained variance. These issues are treated in this chapter.

4.1 ANALYSIS STRATEGY

The number of parameters in a multilevel regression model can easily become very large. If there are p explanatory variables at the lowest level and q explanatory variables at the highest level, the multilevel regression model for two levels is given by equation 4.1:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + \gamma_{pq} Z_{qj} X_{pij} + u_{pj} X_{pij} + u_{0j} + e_{ij}. \quad (4.1)$$

The number of estimated parameters in the model described by equation 4.1 is given by the following list:

Parameters	*Number*
Intercept	1
Lowest-level error variance	1
Fixed slopes for the lowest-level predictors	p
Highest-level error variance	1
Highest-level error variances for these slopes	p
Highest-level covariances of the intercept with all slopes	p
Highest-level covariances between all slopes	$p(p-1)/2$
Fixed slopes for the highest-level predictors	q
Fixed slopes for cross-level interactions	$p \times q$

An ordinary single-level regression model for the same data would estimate only the intercept, one error variance, and $p + q$ regression slopes. The superiority of the multilevel regression model is clear, if we consider that the data are clustered in groups. If we have 100 groups, estimating an ordinary multiple regression model in each group separately requires estimating 100 × (1 regression intercept + 1 residual variance + p regression slopes) plus possible interactions with the q group-level variables. Multilevel regression replaces estimating 100 intercepts by estimating an average intercept plus its residual variance across groups, assuming a normal distribution for these residuals. Thus, multilevel regression analysis replaces estimating 100 separate intercepts by estimating two parameters (the mean and variance of the intercepts), plus a normality assumption. The same simplification is used for the regression slopes. Instead of estimating 100 slopes for the explanatory variable pupil gender, we estimate the average slope along with its variance across groups, and assume that the distribution of the slopes is normal. Nevertheless, even with a modest number of explanatory variables, multilevel regression analysis implies a complicated model. Generally, we do not want to estimate the complete model, first because this is likely to get us into computational problems, but also because it is very difficult to interpret such a complex model. We prefer more limited models that include only those parameters that have proven their worth in previous research, or are of special interest for our theoretical problem.

If we have no strong theories, we can use an exploratory procedure to select a model. Model building strategies can be either top-down or bottom-up. The top-down approach starts with a model that includes the maximum number of fixed and random effects that are considered for the model. Typically, this is done in two steps. The first step starts with all the fixed effects and possible interactions in the model, followed by removing insignificant effects. The second step starts with a rich random structure, followed by removal of insignificant effects. This procedure is described by West et al.

(2007). In multilevel modeling, the top-down approach has the disadvantage that it starts with a large and complicated model, which leads to longer computation time and sometimes to convergence problems. In this book, the opposite strategy is mostly used, which is bottom-up: start with a simple model and proceed by adding parameters, which are tested for significance after they have been added. Typically, the procedure starts by building up the fixed part, and follows after with the random part. The advantage of the bottom-up procedure is that it tends to keep the models simple.

It is attractive to start with the simplest possible model, the intercept-only model, and to add the various types of parameters step by step. At each step, we inspect the estimates and standard errors to see which parameters are significant, and how much residual error is left at the distinct levels. Since we have larger sample sizes at the lowest level, it makes sense to build up the model from there. In addition, since fixed parameters are typically estimated with much more precision than random parameters, we start with the fixed regression coefficients, and add variance components at a later stage. The different steps of such a selection procedure are given below.

Step 1:
Analyze a model with no explanatory variables. This model, the *intercept-only model*, is given by the model of equation 2.8, which is repeated here:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}. \tag{4.2}$$

In equation 4.2, γ_{00} is the regression intercept, and u_{0j} and e_{ij} are the usual residuals at the group and the individual level. The intercept-only model is useful because it gives us an estimate of the intraclass correlation ρ:

$$\sigma^2_{u0} \Big/ \left(\sigma^2_{u0} + \sigma^2_e\right), \tag{4.3}$$

where σ^2_{u0} is the variance of the group-level residuals u_{0j}, and σ^2_e is the variance of the individual-level residuals e_{ij}. The intercept-only model also gives us a benchmark value of the deviance, which is a measure of the degree of misfit of the model, and which can be used to compare models as described in Chapter 3.

Step 2:
Analyze a model with all lower-level explanatory variables fixed. This means that the corresponding variance components of the slopes are fixed at zero. This model is written as:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + u_{0j} + e_{ij}, \tag{4.4}$$

where the X_{pij} are the p explanatory variables at the individual level. In this step, we assess the contribution of each individual-level explanatory variable. The significance of each predictor can be tested, and we can assess what changes occur in the first-level and second-level variance terms. If we use the FML estimation method, we can test the improvement of the final model chosen in this step by computing the difference of the deviance of this model and the previous model (the intercept-only model). This difference approximates a chi-square with, as degrees of freedom, the difference in the number of parameters of both models (see 3.1.1). In this case, the degrees of freedom are simply the number of explanatory variables added in step 2.

Step 3:
Add the higher-level explanatory variables:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + u_{0j} + e_{ij} \tag{4.5}$$

where the Z_{qj} are the q explanatory variables at the group level. This model allows us to examine whether the group-level explanatory variables explain between-group variation in the dependent variable. Again, if we use FML estimation, we can use the global chi-square test to formally test the improvement of fit. If there are more than two levels, this step is repeated on a level-by-level basis.

The models in steps 2 and 3 are often denoted as *variance component* models, because they decompose the intercept variance into different variance components for each hierarchical level. In a variance component model, the regression intercept is assumed to vary across the groups, but the regression slopes are assumed fixed. If there are no higher-level explanatory variables, this model is equivalent to a random effects analysis of covariance (ANCOVA); the grouping variable is the usual ANCOVA factor, and the lowest-level explanatory variables are the covariates (see Kreft & de Leeuw, 1998, p. 30; Raudenbush & Bryk, 2002, p. 25). There is a difference in estimation method: ANCOVA uses OLS techniques and multilevel regression uses ML estimation. Nevertheless, both models are highly similar, and if the groups have all equal sizes, the model is equivalent to analysis of covariance. It is even possible to compute the usual ANCOVA statistics from the multilevel program output (Raudenbush, 1993a). The reason to start with models that include only fixed regression coefficients is that we generally have more information on these coefficients; they can be estimated with more precision than the variance components. When we are confident that we have a well-fitting model for the fixed part, we turn to modeling the random part.

Step 4:
Assess whether any of the slopes of any of the explanatory variables has a significant variance component between the groups. This model, the *random coefficient model*, is given by:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + u_{pj} X_{pij} + u_{0j} + e_{ij} \tag{4.6}$$

where the u_{pj} are the group-level residuals of the slopes of the individual-level explanatory variables X_{pij}.

Testing for random slope variation is best done on a variable-by-variable basis. When we start by including all possible variance terms in a model (which involves also adding many covariance terms), the result is most likely an overparameterized model with serious estimation problems, such as convergence problems or extremely slow computations. Variables that were omitted in step 2 may be analyzed again in this step; it is quite possible for an explanatory variable to have no significant average regression slope (as tested in step 2), but to have a significant variance component for this slope.

After deciding which of the slopes have a significant variance between groups, preferably using the deviance difference test, we add all these variance components simultaneously in a final model, and use the chi-square test based on the deviances to test whether the final model of step 4 fits better than the final model of step 3. Since we are now introducing changes in the random part of the model, the chi-square test can also be used with RML estimation (see 3.1.1). When counting the number of parameters added, remember that adding slope variances in step 4 also adds the covariances between the slopes!

If there are more than two levels, this step is repeated on a level-by-level basis.

Step 5:
Add cross-level interactions between explanatory group-level variables and those individual-level explanatory variables that had significant slope variation in step 4. This leads to the full model:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j + u_{1j} X_{1ij} + u_{0j} + e_{ij}. \tag{4.7}$$

Again, if we use FML estimation, we can use the global chi-square test to formally test the improvement of fit.

If we use an exploratory procedure to arrive at a 'good' model, there is always the possibility that some decisions that have led to this model are based on chance. We may end up overfitting the model by following peculiarities of our specific sample, rather than characteristics of the population. If the sample is large enough, a good strategy is to split it at random into two, then to use one half for our model exploration and the other half for cross-validation of the final model. See Camstra and Boomsma (1992) for a review of several cross-validation strategies. If the sample is not large enough to permit splitting it up in an exploration and validation sample, we can apply a Bonferroni correction to the individual tests performed in the fixed part at each step.

The Bonferroni correction multiplies each p-value by the number of tests performed, and requires the inflated p-value to be significant at the usual level.[1]

At each step, we decide which regression coefficients or (co)variances to keep on the basis of the significance tests, the change in the deviance, and changes in the variance components. Specifically, if we introduce explanatory variables in step 2, we expect the lowest-level variance σ_e^2 to go down. If the composition of the groups with respect to the explanatory variables is not exactly identical for all groups, we expect the higher-level variance σ_{u0}^2 also to go down. Thus, the individual-level explanatory variables explain part of the individual and part of the group variance. The higher-level explanatory variables added in step 3 can explain only group-level variance. It is tempting to compute the analogue of a multiple correlation coefficient to indicate how much variance is actually explained at each level (see Raudenbush & Bryk, 2002). However, this 'multiple correlation' is at best an approximation, and it is quite possible for it to become smaller when we add explanatory variables, which is impossible with a real multiple correlation. This problem is taken up in section 4.5.

4.2 CENTERING AND STANDARDIZING EXPLANATORY VARIABLES

In ordinary multiple regression analysis, linear transformations of the explanatory variables do not change the essence of the regression estimates. If we divide an explanatory variable by two, its new regression slope equals the old one multiplied by two, the standard error is also multiplied by two, and a significance test for the regression slope gives exactly the same result. Most importantly, the proportion of unexplained residual variance, and hence the multiple correlation, does not change either. This is summed up in the statement that the multiple regression model is invariant under linear transformations; if we transform the variables, the estimated parameters change in a similar way, and it is always possible to recalculate the untransformed estimates.

In multilevel regression analysis, the model is only invariant for linear transformations if there are no random regression slopes, that is, if the slopes do not vary across the groups. To understand why this is the case, consider first a simple data set with only one explanatory variable and three groups. Figure 4.1 plots the three

[1] The usual Bonferroni correction is to keep the p-values, and divide the formal alpha level by the number of tests. However, if we have many tests in various steps, we end up with many different significance criteria. It is simpler to correct by appropriately inflating the p-values, and use one alpha criterion for all analysis steps. Both procedures are equivalent, but inflating the p-values makes for a simpler presentation of the results. Holm (1979) describes a more powerful variation of the Bonferroni. If k tests are performed, the Holm correction would multiply the smallest p-value by k, the next smallest p-value by $k - 1$, and so on.

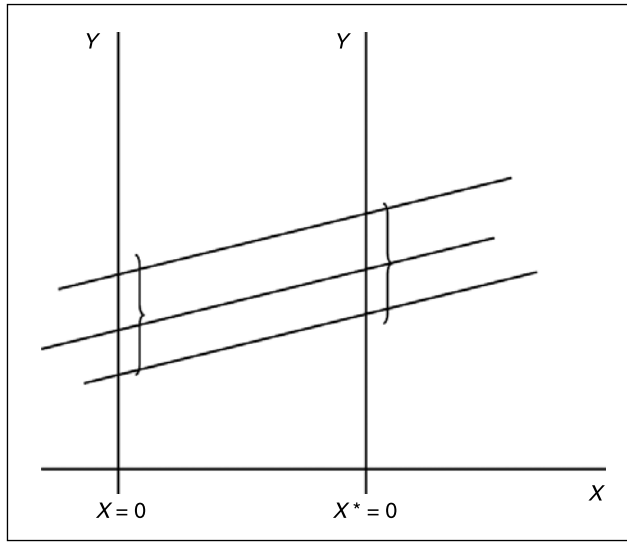

Figure 4.1 Parallel regression lines, with shift on X.

regression slopes when there is no slope variance across the groups. In this situation, the slopes are parallel lines. The variance of the intercept is the variance of the slopes at the point where the slopes cut through the Y-axis, which is at the point where the explanatory variable X is equal to zero. It is clear from Figure 4.1 that, if we shift the scale of X to X^* by adding or subtracting a certain amount, we merely shift the location of the Y-axis, without altering the spread of the intercepts. In this case, the variance of the slope is clearly invariant for shifts on the X-axis, which we produce by adding or subtracting a constant from X.

The variance of the regression slopes is not invariant for such a transformation if the regression slopes vary across the groups, which is the case if we have group-level slope variance. Figure 4.2 shows the situation with three different slope coefficients. This figure shows that with random regression slopes the variance of the intercept changes when we shift the scale of the explanatory variables. It also makes clear how the intercept variance should be interpreted: it is the variance of the intercepts at the point where the explanatory variable X is equal to zero.

It is clear from Figure 4.2 that, if we shift the scale of X to X^* or X^{**} by adding or subtracting a certain amount, the spread of the intercepts changes. If we shift the scale of the X-axis to X^*, the variation of the intercepts is considerable. If we shift the scale of the X-axis to X^{**} and extrapolate the regression slopes, the variation of the intercepts is very small and probably not statistically significant.

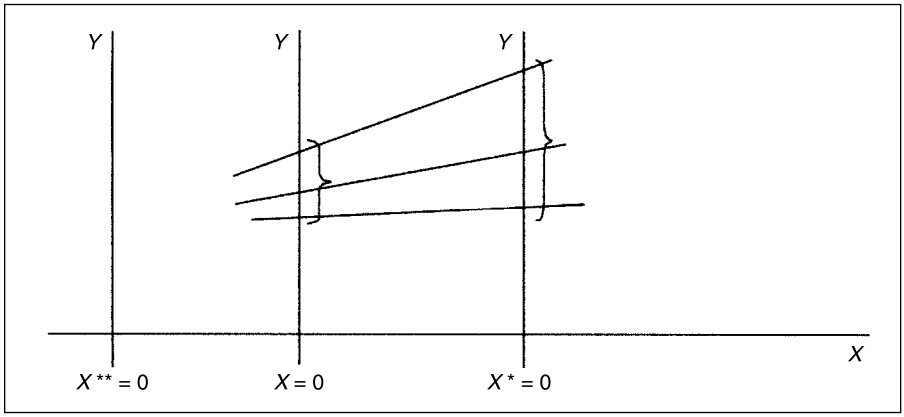

Figure 4.2 Varying regression lines, with shifts on *X*.

In multiple regression analysis, multilevel or single-level, the intercept is interpreted as the expected value of the outcome variable, when all explanatory variables have the value zero. The problem, illustrated in Figure 4.2 by the transformation of the *X*-scale to *X***, is that in many cases 'zero' may not even be a possible value. For instance, if we have an explanatory variable 'gender' coded as 1 = male, 2 = female, zero is not in the possible score range, and as a consequence the value of the intercept is meaningless. To handle this problem, it is useful to perform a transformation of the *X*-variables that make 'zero' a legitimate, observable value.

A linear transformation that is often applied to achieve this is *centering* the explanatory variables. The usual practice is that the overall or grand mean is subtracted from all values of a variable, which is called centering on the grand mean, or grand mean centering in short. If we apply grand mean centering, we solve the problem, because now the intercept in the regression equation is always interpretable as the expected value of the outcome variable, when all explanatory variables have their mean value. Grand mean centering is most often used, but it is also possible to center on a different value, for example on the median, or on a theoretically interesting value. For example, if we have an explanatory variable 'gender' coded as 1 = male, 2 = female, the value of the mean could be 1.6, reflecting that we have 60% females in our sample and 40% males. We can center on the sample mean of 1.6, but we may prefer to center on the mean of a theoretical population with 50% males and 50% females. To accomplish this, we would center on the population mean of 1.5, which effectively leads to a code of –0.5 for male and +0.5 for female. Although we center on a value that as such does not and even cannot exist in the population, the intercept is still interpretable as the expected outcome for the average person, disregarding gender.

In multilevel modeling, centering the explanatory variables has the additional advantage that variances of the intercept and the slopes now have a clear interpretation. They are the expected variances when all explanatory variables are equal to zero, in other words, the expected variances for the 'average' subject.

Centering is also important if the multiple regression model includes interactions. For each of the two explanatory variables involved in an interaction, the interpretation of its slope is that it is the expected value of the slope when the other variable has the value zero. Again, since 'zero' may not even be a possible value, the value of the slope may not be interpretable at all. Since many multilevel regression models include cross-level interactions, this is a serious interpretation problem. When both variables in the interaction are centered on their grand mean, the problem disappears. The problem of interpreting interactions in multilevel regression models is discussed in more detail in the next section.

We will consider the issue of centering in a simple example, using the data from our example in Chapter 2, and including only pupil extraversion as an explanatory variable. We compare the estimates for pupil extraversion as a raw and as a grand mean centered variable, for a random coefficient model (varying slope for pupil extraversion). Table 4.1 also shows the estimates that result when we standardize the variable pupil extraversion. Standardization is a linear transformation that implies grand mean centering, but adds a multiplicative transformation to achieve a standard deviation of 1.

Table 4.1 Popularity predicted by pupil extraversion, raw and centered predictor

Model	Extraversion slope random		
	Raw coeff. (s.e.)	Centered coeff. (s.e.)	Standardized coeff. (s.e.)
Fixed part			
Intercept	2.46 (.20)	5.03 (.10)	5.03 (.10)
Extraversion	0.49 (.03)	0.49 (.03)	0.62 (.03)
Random part			
σ_e^2	0.90 (.03)	0.90 (.03)	0.90 (.03)
σ_{u0}^2	2.94 (.58)	0.88 (.13)	0.88 (.13)
σ_{u2}^2	0.03 (.009)	0.03 (.009)	0.04 (.01)
Deviance	5770.7	5770.7	5770.7

As Table 4.1 shows, grand mean centering of the variable pupil extraversion produces a different estimate for the intercept variance at the second level. The

deviance remains the same, which indicates that all three random coefficient models fit the data equally well. In fact, all three models are equivalent. Equivalent models have the same fit, and produce the same residuals. The parameter estimates are not all identical, but the estimates for one model can be transformed to the estimates of the other model. Thus, grand mean centering and overall standardization do not really complicate the interpretation of the results. In addition, grand mean centering and standardization do have some advantages. One advantage is that the intercept becomes a meaningful value. The value of the higher-level intercept variance also becomes meaningful; it is the expected variance at the mean of all explanatory variables. A second advantage is that with centered explanatory variables the calculations tend to go faster, and encounter fewer convergence problems. Especially when explanatory variables vary widely in their means and variances, grand mean centering or standardization may be necessary to reach convergence, or even to be able to start the computations at all. Since grand mean centering only affects the intercept, which is often not interpreted anyway, it is preferred above standardization, which will also affect the interpretation of the regression slopes and the residual variances.

Some multilevel analysts advocate a totally different way of centering, called group mean centering. Group mean centering means that the group means are subtracted from the corresponding individual scores. Since different values are subtracted from different scores, this is not the same as centering on some overall value, such as the grand mean. This form of centering will be discussed in section 4.4.

4.3 INTERPRETING INTERACTIONS

Whenever there are interactions in a multiple regression analysis (whether these are a cross-level interaction in a multilevel regression analysis or an interaction in an ordinary single-level regression analysis does not matter), there are two important technical points to be made. Both stem from the methodological principle that in the presence of a significant interaction the effect of the interaction variable and the direct effects of the explanatory variables that make up that interaction must be interpreted together as a system (Aiken & West, 1991; Jaccard, Turrisi, & Wan, 1990).

The first point is that if the interaction is significant, it is best to include both direct effects in the regression too, even if they are not significant.

The second point is that in a model with an interaction effect, the regression coefficients of the simple or direct variables that make up that interaction carry a different meaning than in a model without this interaction effect. If there is an interaction, then the regression coefficient of one of the direct variables is the expected value of that regression slope when the other variable is equal to zero, and vice versa. If for one of the variables the value ‘zero’ is widely beyond the range of values that have

been observed, as in age varying from 18 to 55, or if the value 'zero' is in fact impossible, as in gender coded male = 1, female = 2, the result is that the regression coefficient for the other variable has no substantive interpretation. In many such cases, if we compare different models, the regression coefficient for at least one of the variables making up the interaction will be very different from the corresponding coefficient in the model without interaction. *But this change does not mean anything*. One remedy is to take care that the value 'zero' is meaningful and actually occurs in the data. One can accomplish this by centering both explanatory variables on their grand mean.[2] After centering, the value 'zero' refers to the mean of the centered variable, and the regression coefficients do not change when the interaction is added to the model. The regression coefficient of one of the variables in an interaction can now be interpreted as the regression coefficient for individuals with an 'average' score on the other variable. If all explanatory variables are centered, the intercept is equal to the grand mean of the dependent variable.

To interpret an interaction, it is helpful to write out the regression equation for one explanatory variable for various values of the other explanatory variable. The other explanatory variables can be disregarded or included at the mean value. When both explanatory variables are continuous, we write out the regression equation for the lower-level explanatory variable, for a choice of values for the explanatory variable at the higher level. Good choices are the mean and the mean plus/minus 1 standard deviation, or the median and the 25th and 75th percentiles. A plot of these regression lines clarifies the meaning of the interaction. If one of the explanatory variables is dichotomous, we write the regression equation for the continuous variable, for both values of the dichotomous variable.

In the example we have used so far, there is a cross-level interaction between pupil extraversion and teacher experience. In the corresponding data file, pupil extraversion is measured on a 10-point scale, and the range is 1–10. Teacher experience is recorded in years, with the amount of experience ranging from 2 to 25 years. There are no pupils with a zero score on extraversion, and there are no teachers with zero experience, and this explains why adding the cross-level interaction between pupil extraversion and teacher experience to the model results in an appreciable change in the regression slope of pupil extraversion from 0.84 to 1.33. In the model without the interaction, the estimated value for the regression slope of pupil extraversion is independent from teacher experience. Therefore, it can be said to apply to the average class, with an average teacher having an amount of experience somewhere

[2] Standardizing the explanatory variables has the same effect. In this case, it is recommended not to standardize the interaction variable because that makes it difficult to compute predictions or plot interactions. Standardized regression weights for the interaction term can always be determined using equation 2.13.

in the middle between 2 and 25 years. In the model with the interaction, the pupil extraversion slope now refers to a class with a teacher who has zero years of experience. This is an extreme value, which is not even observed in the data. Following the same reasoning, we can conclude that the teacher experience slope refers to pupil extraversion = 0.

The example also includes a variable gender, coded 0 (boys) / 1 (girls). Since 'zero' is a value that does occur in the data, the interpretation of interaction effects with gender is straightforward; leaving the dummy uncentered implies that all slopes for variables interacting with gender refer to boys. This may be awkward in the interpretation, and therefore the dummy variable gender may also be centered around its grand mean or by using effect coding which codes boys = –0.5 and girls = +0.5. Centering issues do not differ for continuous and dichotomous variables (Enders & Tofighi, 2007). For a discussion of different coding schemes for categorical variables see Appendix C of this book.

The estimates for the centered explanatory variables in Table 4.2 are much more comparable across different models than the estimates for the uncentered variables (the small difference between 0.09 and 0.10 for teacher experience is because of rounding). To interpret the cross-level interaction, it helps to work out the regression equations for the effect of pupil extraversion for different values of teacher experience. Using the

Table 4.2 Model without and with cross-level interaction

Model	M1A: main effects	M2: with interaction	M1A: centered interaction variables	M2: centered interaction variables
Fixed part	Coeff. (s.e.)	Coeff. (s.e.)	Coeff. (s.e.)	Coeff. (s.e.)
Intercept	0.74 (.20)	–1.21 (.27)	4.39 (.06)	4.37 (.06)
Gender	1.25 (.04)	1.24 (.04)	1.25 (.04)	1.24 (.04)
Extraversion	0.45 (.02)	0.80 (.04)	0.45 (.02)	0.45 (.02)
T. exp	0.09 (.01)	0.23 (.02)	0.09 (.01)	0.10 (.01)
Extra × T.exp		–0.03 (.003)		–0.025 (.002)
Random part				
σ^2_e	0.55 (.02)	0.55 (.02)	0.55 (.02)	0.55 (.02)
σ^2_{u0}	1.28 (.28)	0.45 (.16)	0.28 (.04)	0.28 (.04)
σ^2_{u2}	0.03 (.008)	0.005 (.004)	0.03 (.008)	0.005 (.004)
$\sigma_{u_{02}}$	–0.18 (.05)	–0.03 (.02)	–0.01 (.02)	–0.00 (.01)
Deviance	4812.8	4747.6	4812.8	4747.6

centered variables, the regression equation for the effect of pupil extraversion on popularity is:

$$popularity = 4.368 + 1.241 \times gender + 0.451 \times extrav + 0.097 \times t.exp - 0.025 \times t.exp \times extrav.$$

The average effect of a change of one scale point in extraversion is to increase the popularity by 0.451. This is the predicted value for teachers of average experience (14.2 years in the raw data set, zero in the centered data). For each year more, the effect of extraversion decreases by 0.025. So for the teachers with the most experience, 25 years in our data, the expected effect of extraversion is $0.451 - 0.025 \times (25 - 14.2) = 0.18$. So, for these teachers the effect of extraversion is predicted to be much smaller.

Another way to make it easier to interpret an interaction is to plot the regression slopes for one of the explanatory variables for some values of the other. The mean of pupil gender is 0.51, so we can absorb that into the intercept, giving:

$$popularity = 5.001 + 0.451 \times extrav + 0.097 \times t.exp - 0.025 \times t.exp \times extrav.$$

The centered variable pupil extraversion ranges from –4.22 to 4.79. The centered variable teacher experience ranges from –12.26 to 10.74, with a standard deviation of 6.552. We can use equation 2.12 to predict popularity, with extraversion ranging from –4.22 to 4.79 and teacher experience set at –6.552, 0, and 6.552, which are the values of 1 standard deviation below the mean, the mean, and 1 standard deviation above the mean. Figure 4.3 presents a plot of the three regression lines.

It is clear that more extraverted pupils have a higher expected popularity score, and that the difference is smaller with more experienced teachers. In general, more experienced teachers have classes with a higher average popularity. At the maximum values of teacher experience, this relationship appears to reverse, but these differences are probably not significant. If we used the uncentered scores for the plot, the scale of the *X*-axis, which represents teacher experience, would change, but the picture would not. Centering explanatory variables is especially attractive when we want to interpret the meaning of an interaction by inspecting the regression coefficients in Table 4.2. As Figure 4.3 shows, plotting the interaction over the range of observed values of the explanatory variables is an effective way to convey its meaning, even if we work with raw variables.

Interactions are sometimes interpreted in terms of moderation or a moderator effect. In Figure 4.3, we can state that the effect of pupil extraversion is moderated by teacher experience, or that the effect of teacher experience is moderated by pupil extraversion. In multilevel analysis, where the interest often lies in contextual effects, the interpretation that the effect of pupil extraversion is moderated by teacher

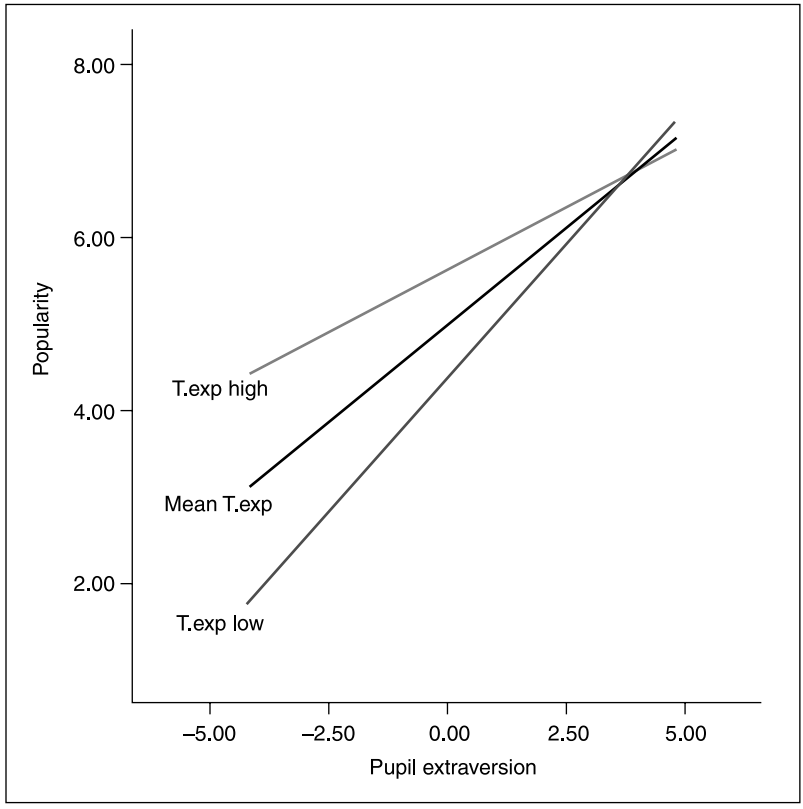

Figure 4.3 Regression lines for popularity by extraversion, for three levels of teacher experience.

experience would in many cases be preferred. A more statistical approach is to examine the range of values of teacher experience for which the effect of pupil extraversion is significant. A simple approach to probing interactions is to test simple slopes at specific levels of the predictors. In this approach, teacher experience is centered on a range of values, to find out by trial and error where the boundary lies between a significant and a nonsignificant effect for pupil extraversion. A more general method is the Johnson-Neyman (J-N) technique, which views interactions as conditional relations in which the effect of one predictor varies with the value of another. The values of the regression coefficients and their standard errors are used to calculate the range of values on one explanaory variable for which the other variable in the interaction shows a significant effect. Bauer and Curran (2005) describe these techniques in the context of standard and multilevel regression analysis, and Preacher, Curran, and Bauer (2006)